PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12N 15/90, 5/16, A01K 67/027, C07K 16/18	A1	(11) International Publication Number: WO 98/40510 (43) International Publication Date: 17 September 1998 (17.09.98)

(21) International Application Number: PCT/US98/04687

(22) International Filing Date: 11 March 1998 (11.03.98)

(30) Priority Data:

60/040,664	11 March 1997 (11.03.97)	US
60/053,868	28 July 1997 (28.07.97)	US
60/065,303	13 November 1997 (13.11.97)	US

(71) Applicant *(for all designated States except US)*: REGENTS OF THE UNIVERSITY OF MINNESOTA [US/US]; 100 Church Street, S.E., Minneapolis, MN 55455 (US).

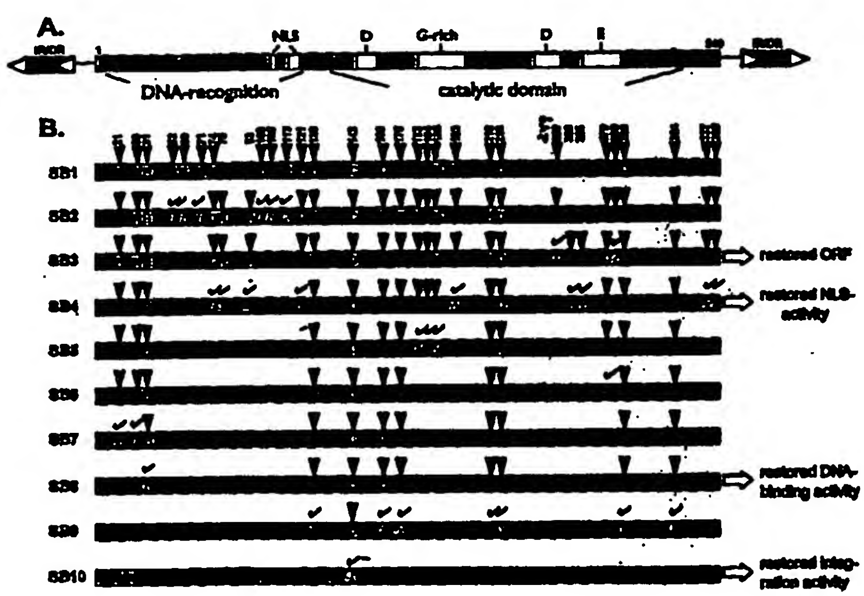
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: HACKETT, Perry, B. [US/US]; 4971 Virginia, Shoreview, MN 55126 (US). IVICS, Zoltan [HU/HU]; Netherlands Cancer Institute, Division of Molecular Biology, Plesmanlaan 121, NL-1066 CX Amsterdam (NL). IZSVAK, Zsuzsanna [HU/HU]; Netherlands Cancer Institute, Division of Molecular Biology, Plesmanlaan 121, NL-1066 CX Amsterdam (NL). CALDOVIC, Ljubica [YU/YU]; 250 Bio-Sciene Center, University of Minnesota, 1445 Gortner Avenue, St. Paul, MN 55108-1095 (US).

(74) Agent: BUHARIN, Amelia, A.; Mueting, Raasch & Gebhardt, P.A., P.O. Box 581415, Minneapolis, MN 55458-1415 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: DNA-BASED TRANSPOSON SYSTEM FOR THE INTRODUCTION OF NUCLEIC ACID INTO DNA OF A CELL



(57) Abstract

This invention relates to a system for introducing nucleic acid into the DNA of a cell. The system includes the use of a member of the SB family of transposases (SB) or nucleic acid encoding the transposase and a nucleic acid fragment that includes a nucleic acid sequence with flanking inverted repeats. The transposase recognizes at least a portion of an inverted repeats and incorporates the nucleic acid sequence into the DNA. Methods for use of this system are discussed.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

DNA-BASED TRANSPOSON SYSTEM FOR THE INTRODUCTION OF NUCLEIC ACID INTO DNA OF A CELL

Field of the Invention

This invention relates to methods for gene expression, mapping genes, mutagenesis, methods for introducing DNA into a host chromosome and to transposons and transposases.

Transposons or transposable elements include a short piece of nucleic acid bounded by repeat sequences. Active transposons encode enzymes that facilitate the insertion of the nucleic acid into DNA sequences.

In vertebrates, the discovery of DNA-transposons, mobile elements that move via a DNA intermediate, is relatively recent (Radice, A.D., et al., 1994. *Mol. Gen. Genet.* 244, 606-612). Since then, inactive, highly mutated members of the Tc1/*mariner* as well as the hAT (*hobo/Ac/Tam*) superfamilies of eukaryotic transposons have been isolated from different fish species, *Xenopus* and human genomes (Oosumi et al., 1995. *Nature* 378, 873; Ivics et al. 1995. *Mol. Gen. Genet.* 247, 312-322; Koga et al., 1996. *Nature* 383, 30; Lam et al., 1996. *J. Mol. Biol.* 257, 359-366 and Lam, W. L., et al. *Proc. Natl. Acad. Sci. USA* 93, 10870-10875).

These transposable elements transpose through a cut-and-paste mechanism; the element-encoded transposase catalyzes the excision of the transposon from its original location and promotes its reintegration elsewhere in the genome (Plasterk, 1996 *Curr. Top. Microbiol. Immunol.* 204, 125-143). Autonomous members of a transposon family can express an active transposase, the *trans*-acting factor for transposition, and thus are capable of transposing on their own. Nonautonomous elements have mutated transposase genes but may retain *cis*-acting DNA sequences. These cis-acting DNA sequences are also referred to as inverted terminal repeats. Some inverted repeat sequences include one or more direct repeat sequences. These sequences usually are embedded in the terminal inverted repeats (IRs) of the elements, which are required for mobilization in the presence of a complementary transposase from another element.

Not a single autonomous element has been isolated from vertebrates; all transposon-like sequences are defective, apparently as a result of a process called "vertical inactivation" (Lohe et al., 1995 *Mol. Biol. Evol.* 12, 62-72). According to one phylogenetic model (Hartl et al., 1997 *Trends Genet.* 13, 197-201), the ratio of nonautonomous to autonomous elements in eukaryotic genomes increases as a result of the *trans*-complementary nature of transposition. This process leads to a state where the ultimate disappearance of active, transposase-producing copies in a genome is inevitable. Consequently, DNA-transposons can be viewed as transitory components of genomes which, in order to avoid extinction, must find ways to establish themselves in a new host. Indeed, horizontal gene transmission between species is thought to be one of the important processes in the evolution of transposons (Lohe et al., 1995 *supra* and Kidwell, 1992. *Curr. Opin. Genet. Dev.* 2, 868-873).

The natural process of horizontal gene transfer can be mimicked under laboratory conditions. In plants, transposable elements of the *Ac/Ds* and *Spm* families have been routinely introduced into heterologous species (Osborne and Baker, 1995 *Curr. Opin. Cell Biol.* 7, 406-413). In animals, however, a major obstacle to the transfer of an active transposon system from one species to another has been that of species-specificity of transposition due to the requirement for factors produced by the natural host. For this reason, attempts have been unsuccessful to use the P element transposon of *Drosophila melanogaster* for genetic transformation of non-drosophilid insects, zebrafish and mammalian cells (Gibbs et al., 1994 *Mol. Mar. Biol. Biotech.* 3, 317-326; Handler et al., 1993. *Arch. Insect Biochem. Physiol.* 22, 373-384; and Rio et al., 1988 *J. Mol. Biol.* 200, 411-415). In contrast to P elements, members of the Tc1/*mariner* superfamily of transposable elements may not be as demanding for species-specific factors for their transposition. These elements are widespread in nature, ranging from single-cellular organisms to humans (Plasterk, 1996, *supra*). In addition, recombinant Tc1 and *mariner* transposases expressed in *E. coli* are sufficient to catalyze transposition *in vitro* (Vos et al, 1996 *Genes. Dev.* 10, 755-761 and Lampe et al., 1996. *EMBO J.* 15, 5470-5479 and PCT International Publication No. WO 97/29202 to Plasterk et al.). Furthermore,

gene vectors based on *Minos*, a Tc1-like element (TcE) endogenous to *Drosophila hydei*, were successfully used for germline transformation of the fly *Ceratitis capitata* (Loukeris et al., 1995 *Science* 270, 2002-2005).

Molecular phylogenetic analyses have shown that the majority of the fish TcEs can be classified into three major types: zebrafish-, salmonid- and *Xenopus* TXr-type elements, of which the salmonid subfamily is probably the youngest and thus most recently active (Ivics et al., 1996, *Proc. Natl. Acad. Sci. USA* 93, 5008-5013). In addition, examination of the phylogeny of salmonid TcEs and that of their host species provides important clues about the ability of this particular subfamily of elements to invade and establish permanent residences in naive genomes through horizontal transfer, even over relatively large evolutionary distances.

TcEs from teleost fish (Goodier and Davidson, 1994 *J. Mol. Biol.* 241, 26-34 and Izsvak et al., 1995. *Mol. Gen. Genet.* 247, 312-322), including Tdr1 in zebrafish (Izsvak et al., 1995, *supra*) and other closely related TcEs from nine additional fish species (Ivics et al., 1996. *Proc. Natl. Acad. Sci. USA* 93, 5008-5013) are by far the best characterized of all the DNA-transposons known in vertebrates. Fish elements, and other TcEs in general, are typified by a single gene encoding a transposase enzyme flanked by inverted repeat sequences. Unfortunately, all the fish elements isolated so far are inactive due to one or more mutations in the transposase genes.

Methods for introducing DNA into a cell are known. These include, but are not limited to, DNA condensing reagents such as calcium phosphate, polyethylene glycol, and the like), lipid-containing reagents, such as liposomes, multi-lamellar vesicles, and the like, and virus-mediated strategies. These methods all have their limitations. For example, there are size constraints associated with DNA condensing reagents and virus-mediated strategies. Further, the amount of nucleic acid that can be introduced into a cell is limited in virus strategies. Not all methods facilitate integration of the delivered nucleic acid into cellular nucleic acid and while DNA condensing methods and lipid-containing reagents are relatively easy to prepare, the incorporation of nucleic acid into viral vectors can be labor intensive. Moreover, virus-mediated

strategies can be cell-type or tissue-type specific and the use of virus-mediated strategies can create immunologic problems when used *in vivo*.

There remains a need for new methods for introducing DNA into a cell, particularly methods that promote the efficient integration of nucleic acid fragments of varying sizes into the nucleic acid of a cell, particularly the integration of DNA into the genome of a cell.

Summary of the Invention

We have developed a DNA-based transposon system for genome manipulation in vertebrates. Members of the Tc1/*mariner* superfamily of transposons are prevalent components of the genomes of teleost fish as well as a variety of other vertebrates. However, all the elements isolated from nature appear to be transpositionally inactive. Molecular phylogenetic data were used to identify a family of synthetic, salmonid-type Tc1-like transposases (SB) with their recognition sites that facilitate transposition. A consensus sequence of a putative transposase gene was first derived from inactive elements of the salmonid subfamily of elements from eight species of fish and then engineered by eliminating the mutations that rendered these elements inactive. A transposase was created in which functional domains were identified and tested for biochemical functions individually as well as in the context of a full-length transposase. The transposase binds to two binding-sites within the inverted repeats of salmonid elements, and appears to be substrate-specific, which could prevent cross-mobilization between closely related subfamilies of fish elements. *SB* transposases significantly enhance chromosomal integration of engineered transposons not only in fish, but also in mouse and in human cells. The requirements for specific motifs in the transposase plus specific sequences in the target transposon, along with activity in fish and mammalian cells alike, establishes *SB* transposase as the first active DNA-transposon system for germline transformation and insertional mutagenesis in vertebrates.

In one aspect of this invention, the invention relates to a nucleic acid fragment comprising: a nucleic acid sequence positioned between at least two inverted repeats wherein the inverted repeats can bind to a SB protein and

wherein the nucleic acid fragment is capable of integrating into DNA in a cell. In one embodiment nucleic acid fragment is part of a plasmid and preferably the nucleic acid sequence comprises at least a portion of an open reading frame and also preferably at least one expression control region of a gene. In one embodiment, the expression control region is selected from the group consisting of a promoter, an enhancer or a silencer. Preferably the nucleic acid sequence comprises a promoter operably linked to at least a portion of an open reading frame.

In one embodiment the cell is obtained from an animal such as an invertebrate or a vertebrate. Preferred invertebrates include crustacean or a mollusk including, but not limited to a shrimp, a scallop, a lobster, a clam or an oyster. Preferred vertebrate embodiments include fish, birds, and mammal such as those selected from the group consisting of mice, ungulates, sheep, swine, and humans. The DNA of the cell can be the cell genome or extrachromosomal DNA, including an episome or a plasmid.

In one embodiment of this aspect of the invention, at least one of the inverted repeats comprises SEQ ID NO:4 or SEQ ID NO: 5 and preferably the amino acid sequence of the SB protein has at least an 80% amino acid identity to SEQ ID NO: 1. Also preferably, at least one of the inverted repeats comprises at least one direct repeat, wherein the at least one direct repeat sequence comprises SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 8 or SEQ ID NO:9. A preferred direct repeat is SEQ ID NO:10. Also preferably the nucleic acid fragment includes a direct repeat that has at least an 80% nucleic acid sequence identity to SEQ ID NO: 10.

In another aspect of this invention, the invention relates to a gene transfer system to introduce DNA into the DNA of a cell comprising: a nucleic acid fragment comprising a nucleic acid sequence positioned between at least two inverted repeats wherein the inverted repeats can bind to an SB protein and wherein the nucleic acid fragment is capable of integrating into DNA of a cell; and a transposase or nucleic acid encoding a transposase, wherein the transposase is an SB protein with an amino acid sequence sharing at least an

80% identity to SEQ ID NO:1. In one embodiment, the the SB protein comprises SEQ ID NO:1. Alternatively, the SB protein is encoded by DNA that can hybridize to SEQ ID NO: 1under stringent hybridization conditions. In one embodiment, the transposase is provided to the cell as a protein and in another the transposase is provided to the cell as nucleic acid. In one embodiment the nucleic acid is RNA and in another the nucleic acid is DNA. In yet another embodiment, the nucleic acid encoding the transposase is integrated into the genome of the cell. The nucleic acid fragment can be part of a plasmid or a recombinant viral vector. Preferably, the nucleic acid sequence comprises at least a portion of an open reading frame and also preferably, the nucleic acid sequence comprises at least a regulatory region of a gene. In one embodiment the regulatory region is a transcriptional regulatory region and the regulatory region is selected from the group consisting of a promoter, an enhancer, a silencer, a locus-control region, and a border element. In another embodiment, the nucleic acid sequence comprises a promoter operably linked to at least a portion of an open reading frame.

The cells used in this aspect of the invention can be obtained from a variety of sources including bacteria, fungi, plants and animals. In one embodiment, the cells are obtained from an animal; either a vertebrate or an invertebrate. Preferred invertebrate cells include crustaceans or a mollusks. Preferred vertebrates include fish, birds, and mammal such as rodents, ungulates, sheep, swine and humans.

The DNA of the cell receiving the nucleic acid fragment can be a part of the cell genome or extrachromosal DNA. Preferably, the inverted repeats of the gene transfer system comprise SEQ ID NO:4 or SEQ ID NO:5. Also preferably the amino acid sequence of the SB protein has at least a 80% identity to SEQ ID NO:1 and preferably at least one of the inverted repeats comprises at least one direct repeat and wherein the at least one direct repeat sequence comprises SEQ ID NO:6, SEQ ID NO: 7, SEQ ID NO:8 or SEQ ID NO:9. In one embodiment, the direct repeat has a consensus sequence of SEQ ID NO:10. In a particularly preferred embodiment, the nucleic acid sequence is part of a library of recombinant sequences and the nucleic acid sequence is introduced into the cell

using a method selected from the group consisting of: particle bombardment, electroporation, microinjection, combining the nucleic acid fragment with lipid-containing vesicles or DNA condensing reagents, and incorporating the nucleic acid fragment into a viral vector and contacting the viral vector with the cell.

In another aspect of this invention, the invention relates to nucleic acid encoding an SB protein, wherein the nucleic acid encodes a protein comprising SEQ ID NO:1 or a protein comprising an amino acid sequence with at least 80% identity to SEQ ID NO:1. The nucleic acid encoding the SB protein can be incorporated into a nucleic acid vector, such as a gene expression vector either as a viral vector or as a plasmid. The nucleic acid can be circular or linear. This invention also relates to cells expessing the SB protein.

In one embodiment the cells containing the SB protein cell are obtained from an animal, either a vertebrate or an invertebrate. Prefered vertebrates include fish, birds and mammals. The cells can be obtained from a variety of tissues including pluripotent and totipotent cells such as an oocyte, one or more cells of an embryo, or an egg. In one embodiment, the cell is part of a tissue or organ. In one embodiment, the nucleic acid encoding the SB protein is integrated in the genome of a cell.

The invention also relates to SB protein comprising the amino acid sequence of SEQ ID NO:1.

In addition, the invention relates to a method for producing a transgenic animal comprising the steps of: introducing a nucleic acid fragment and a transposase into a pluripotent or totipotent cell wherein the nucleic acid fragment comprises a nucleic acid sequence positioned between at least two inverted repeats, wherein the inverted repeats can bind to a SB protein and wherein the nucleic acid fragment is capable of integrating into DNA in a cell and wherein the transposase is an SB protein having an amino acid sequence identity of least 80% to SEQ ID NO:1; and growing the cell into an animal. Preferred pluripotent or totipotent cells include an oocyte, a cell of an embryo, an egg and a stem cell. In one embodiment, the introducing step comprises a method selected from the group consisting of: microinjection; combining the nucleic acid fragment with cationic lipid vesicles or DNA condensing reagents; and

incorporating the nucleic acid fragment into a viral vector and contacting the viral vector with the cell as well as particle bombardment and electroporation. In another preferred embodiment the viral vector is selected from the group consisting of a retroviral vector, an adenovirus vector, a herpesvirus or an adeno-associated viral vector. Preferred animals used in this method include a mouse, a fish, an ungulate, a bird, or a sheep.

In yet another aspect of this invention, the invention relates to a method for introducing nucleic acid into DNA in a cell comprising the step of: introducing a nucleic acid fragment comprising a nucleic acid sequence positioned between at least two inverted repeats into a cell wherein the inverted repeats can bind to an SB protein and wherein the nucleic acid fragment is capable of integrating into DNA in a cell in the presence of an SB protein. In a preferred embodiment, the method further comprises introducing an SB protein into the cell. In one embodiment, the SB protein has an amino acid sequence comprising at least a 80% identity to SEQ ID NO:1. The SB protein can be introduced into the cell as protein or as nucleic acid, including RNA or DNA. The cell receiving the nucleic acid fragment can already include nucleic acid encoding an SB protein and already express the protein. In a one embodiment, the SB protein is integrated into the cell genome. The SB protein can be stably expressed in the cell or transiently expressed and nucleic acid encoding the SB protein can be under the control of an inducible promoter or under the control of a constitutive promoter. In one aspect of this method, the introducing step comprises a method for introducing nucleic acid into a cell selected from the group consisting of: microinjection; combining the nucleic acid fragment with cationic lipid vesicles or DNA condensing reagents; and incorporating the nucleic acid fragment into a viral vector and contacting the viral vector with the cell. Preferred viral vectors are selected from the group consisting of a retroviral vector, an adenovirus vector or an adeno-associated viral vector. In another aspect of this method, the method includes the step of introducing an SB protein or RNA encoding an SB protein into the cell. The cells used for this method can be pluripotent or a totipotent cell and this invention also relates to transgenic animals produced by this method. Where transgenic animals are produced, the

nucleic acid sequence preferably encodes a protein and preferably a protein to be collected from the transgenic animal or a marker protein. The invention also relates to those cells of the transgenic animal expressing the protein encoded by the nucleic acid sequence.

The invention also relates to a SB protein. In one embodiment the protein has the following characteristics: an ability to catalyze the integration of nucleic acid into DNA of a cell; capable of binding to the inverted repeat sequence of SEQ ID NOS 4 or 5; and 80% amino acid sequence identity to SEQ ID NO:1. In another embodiment, the protein has the following characteristics: transposase activity; a molecular weight range of about 35 kD to about 40 kD on about a 10% SDS-polyacrylamide gel; and an NLS sequence, a DNA binding domain and a catalytic domain and wherein the protein has at least about five-fold improvement in the rate for introducing a nucleic acid fragment into the nucleic acid of a cell as compared to the level obtained by non-homologous recombination. Preferred methods for testing the rate of nucleic acid fragment incorporation is provided in the examples.

In yet another aspect, the invention relates to a method for mobilizing a nucleic acid sequence in a cell comprising the steps of: introducing the protein of this invention into a cell housing DNA containing the nucleic acid fragment of this invention, wherein the protein mobilizes the nucleic acid fragment from a first position within the DNA of a cell to a second position within the DNA of the cell. In one embodiment, the DNA of a cell is genomic DNA. In another, the first position within the DNA of a cell is extrachromosomal DNA and in yet another, the second position within the DNA of a cell is extrachromosomal DNA. In a preferred embodiment, the protein is introduced into the cell as RNA.

The invention also relates to a method for identifying a gene in a genome of a cell comprising the steps of: introducing a nucleic acid fragment and an SB protein into a cell, wherein the nucleic acid fragment comprises a nucleic acid sequence positioned between at least two inverted repeats into a cell wherein the inverted repeats can bind to the SB protein and wherein the nucleic acid fragment is capable of integrating into DNA in a cell in the presence of the SB

protein; digesting the DNA of the cell with a restriction endonuclease capable of cleaving the nucleic acid sequence; identifying the inverted repeat sequences; sequencing the nucleic acid close to the inverted repeat sequences to obtain DNA sequence from an open reading frame; and comparing the DNA sequence with sequence information in a computer database. In one embodiment, the restriction endonuclease recognizes a 6-base recognition sequence. In another embodiment, the digesting step further comprises cloning the digested fragments or PCR amplifying the digested fragments.

The invention also relates to a stable transgenic vertebrate line comprising a gene operably linked to a promoter, wherein the gene and promoter are flanked by inverted repeats, wherein the inverted repeats can bind to an SB protein. In one embodiment, the SB protein comprises SEQ ID NO:1 or an amino acid sequence with at least 80% homology to SEQ ID NO:1. In one embodiment, the vertebrate is a fish, including a zebrafish and in another the vertebrate is a mouse.

In addition, the invention also relates to a protein with transposase activity that can bind to one or more of the following sequences: SEQ ID NO: 4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, and SEQ ID NO:10.

Brief Description of the Figures

Fig. 1 illustrates the molecular reconstruction of a salmonid Tc1-like transposase gene. **Fig. 1(A)** is a schematic map of a salmonid TcE with the conserved domains in the transposase and IR/DR (inverted repeat/direct repeat) flanking sequences. **Fig. 1(B)** provides an exemplary strategy for constructing an open reading frame for a salmonid transposase (SB1-SB3) and then systematically introducing amino acid replacements into this gene (SB4-SB10). Amino acid residues are shown using single letter code, typed black when different from the consensus. Positions within the transposase polypeptide that were modified by site-specific mutagenesis are indicated with arrows. Translational termination codons appear as asterisks, frameshift mutations are shown as #. Residues changed to the consensus are check-marked and typed in white italics. In the right margin, the various functional tests that were done at

various stages of the reconstruction are indicated.

Fig. 2(A) is the amino acid sequence (sEQ ID NO:1) of an SB transposase. The major functional domains are highlighted. **Fig. 2(B)** is a nucleic acid sequence encoding the SB protein (SEQ ID NO:3).

Fig. 3 illustrates the DNA-binding activities of an N-terminal derivative of the *SB* transposase. **Fig. 3(A)** provides the SDS-PAGE analysis illustrating the steps in the expression and purification of N123. Lanes: 1) extract of cells containing expression vector pET21a; 2) extract of cells containing expression vector pET21a/N123 before induction with IPTG; 3) extract of cells containing expression vector pET21a/N123 after 2.5 h of induction with IPTG; 4) partially purified N123 using Ni^{2+}-NTA resin. Molecular weights in kDa are indicated on the right. **Fig. 3(B)** illustrates the results of mobility-shift analysis studies to determine whether N123 bound to the inverted repeats of fish transposons. Lanes: 1) probe only; 2) extract of cells containing expression vector pET21a; 3) 10,000-fold dilution of the N123 preparation shown in lane 4 of Panel A; 4) same as lane 3 plus a 1000-fold molar excess of unlabelled probe as competitor DNA; 5) same as lane 3 plus a 1000-fold molar excess of an inverted repeat fragment of a zebrafish Tdr1 element as competitor DNA; 6-13) 200,000-, 100,000-, 50,000-, 20,000-, 10,000-, 5,000-, 2,500-, and 1,000-fold dilutions of the N123 preparation shown in lane 4 of Panel A.

Fig. 4 provides the DNase I footprinting of deoxyribonucleoprotein complexes formed by N123. **Fig. 4(A)** is a photograph of a DNase I footprinting gel containing a 500-fold dilution of the N123 preparation shown in lane 4 of Fig. 3A using the same transposon inverted repeat DNA probe as in Fig. 3B. Reactions were run in the absence (bottom lane) or presence (middle lane) of N123. Maxam-Gilbert sequencing of purine bases in the same DNA was used as a marker (lane 1). Reactions were run in the presence (lane 2) or absence (lane 3) of N123. **Fig 4(B)** provides a sequence comparison of the salmonid transposase-binding sites illustrated in Panel A with the corresponding sequences in the zebrafish Tdr1 elements. **Fig. 4(C)** is a sequence comparison between the outer and internal transposase-binding sites in the SB transposons.

Fig. 5 illustrates the integration activity of *SB* in human HeLa cells. **Fig.**

5(A) is a schematic illustrating the genetic assay strategy for *SB*-mediated transgene integration in cultured cells. **Fig. 5(B)** demonstrates HeLa cell integration using Petri dishes of HeLa cells with stained colonies of G-418-resistant HeLa cells that were transfected with different combinations of donor and helper plasmids. Plate: 1) pT/neo plus pSB10-AS; 2) pT/neo plus pSB10; 3) pT/neo plus pSB10-ΔDDE; 4) pT/neo plus pSB6; 5) pT/neo-ΔIR plus pSB10.

Fig. 6 summarizes the results of transgene integration in human HeLa cells. Integration was dependent on the presence of an active *SB* transposase and a transgene flanked by transposon inverted repeats. Different combinations of the indicated donor and helper plasmids were cotransfected into cultured HeLa cells and one tenth of the cells, as compared to the experiments shown in Fig. 5, were plated under selection to count transformants. The efficiency of transgene integration was scored as the number of transformants surviving antibiotic selection. Numbers of transformants at right represent the numbers of G-418-resistant cell colonies per dish. Each number represents the average obtained from three transfection experiments.

Fig. 7 illustrates the integration of neomycin resistance-marked transposons into the chromosomes of HeLa cells. **Fig. 7(A)** illustrates the results of a southern hybridization of HeLa cell genomic DNA with neomycin-specific radiolabeled probe from 8 individual HeLa cell clones that had been cotransfected with pT/neo and pSB10 and survived G-418 selection. Genomic DNA was digested with the restriction enzymes *Nhe*I, *Xho*I, *Bgl*II, *Spe*I and *Xba*I, enzymes that do not cut within the *neo*-marked transposon, prior to agarose gel electrophoresis and blotting. **Fig. 7(B)** is a diagram of the junction sequences of T/neo transposons integrated into human genomic DNA. The donor site is illustrated on top with plasmid vector sequences that originally flanked the transposon in pT/neo (black arrows). Human genomic DNA serving as target for transposon insertion is illustrated as a gray box. IR sequences are uppercase with flanking sequences in lowercase.

Fig. 8 is a schematic demonstrating an interplasmid assay for excision and integration of a transposon. The assay was used to evaluate transposase activity in zebrafish embryos. Two plasmids plus an RNA encoding an SB

transposase protein were coinjected into the one-cell zebrafish embryo. One of the plasmids had an ampicillin resistance gene (Ap) flanked by IR/DR sequences recognizable by the SB transposase. Five hours after fertilization and injection, low molecular weight DNA was isolated from the embryos and used to transform *E. coli*. The bacteria were grown on media containing ampicillin and kanamycin (Km) to select for bacteria harboring single plasmids containing both the Km and Ap antibiotic-resistance markers. The plasmids from doubly resistant cells were examined to confirm that the Ap-transposon was excised and reintegrated into the Km target plasmid. Ap-transposons that moved into either another indicator Ap-plasmid or into the zebrafish genome were not scored. Because the amount of DNA in injected plasmid was almost equal to that of the genome, the number of integrations of Ap-transposons into target plasmids approximated the number of integrations into the genome.

Fig. 9 illustrates two preferred methods for using the gene transfer system of this invention. Depending on the integration site of the nucleic acid fragment of this invention the effect can be either a loss-of-function or a gain-of-function mutation. Both types of activity can be exploited, for example, for gene discovery and/or functional genomics.

Fig. 10 illustrates a preferred screening strategy using IRS-PCR (interspersed repetitive sequence polymerase chain reaction). **Fig. 10A** illustrates a chromosomal region in the zebrafish genome containing the retroposon DANA (D) 5'-GGCGACRCAGTGGCGCAGTRGG (SEQ ID NO:13) and 5'-GAAYRTGCAAACTCCACACAGA (SEQ ID NO:14); Tdr1 transposons (T) 5'-TCCATCAGACCACAGGACAT (SEQ ID NO:15) and 5'-TGTCAGGAGGAATGGGCCAAAATTC (SEQ ID NO:16); and *Angel* (A) (a highly reiterated miniature inverted-repeat transposable element) 5'-TTTCAGTTTTGGGTGAACTATCC (SEQ ID NO:12) sequences. The arrows above the elements represent specific PCR primers.

The X superimposed on the central DANA element is meant to represent a missing element or a mutated primer binding site in the genome of another zebrafish strain. The various amplified sequence tagged sites (STSs) are

identified by lowercase letter, beginning with the longest detectable PCR product. The products marked with an X are not produced in the PCR reaction if genomes with defective "X-DNA" are amplified. Elements separated by more than 3000 base pairs (bp) and elements having the wrong orientation relative to each other are not amplified efficiently. **Fig. 10B** is a schematic of the two sets of DNA amplification products from both genomes with (lane 1) and without (lane 2) the X'ed DANA element. Note that bands "a" and "d" are missing when the marked DANA sequence is not present.

Detailed Description of the Preferred Embodiments

The present invention relates to novel transposases and the transposons that are used to introduce nucleic acid sequences into the DNA of a cell. A transposase is an enzyme that is capable of binding to DNA at regions of DNA termed inverted repeats. Transposons typically contain at least one, and preferably two, inverted repeats that flank an intervening nucleic acid sequence. The transposase binds to recognition sites in the inverted repeats and catalyzes the incorporation of the transposon into DNA. Inverted repeats of an SB transposon can include two direct repeats and include at least one direct repeat.

Transposons are mobile, in that they can move from one position on DNA to a second position on DNA in the presence of a transposase. There are two fundamental components of any mobile cut-and-paste type transposon system, a source of an active transposase and the DNA sequences that are recognized and mobilized by the transposase. Mobilization of the DNA sequences permits the intervening nucleic acid between the recognized DNA sequences to also be mobilized.

DNA-transposons, including members of the Tc1/*mariner* superfamily, are ancient residents of vertebrate genomes (Radice et al., 1994; Smit and Riggs, 1996 *Proc. Natl. Acad. Sci. USA* 93, 1443-1448). However, neither autonomous copies of this class of transposon nor a single case of a spontaneous mutation caused by a TcE insertion have been proven in vertebrate animals. This is in contrast to retroposons whose phylogenetic histories of mutating genes in vertebrates is documented (Izsvak et al., 1997). Failure to isolate active DNA-

transposons from vertebrates has greatly hindered ambitions to develop these elements as vectors for germline transformation and insertional mutagenesis. However, the apparent capability of salmonid TcEs for horizontal transmission between two teleost orders (Ivics et al., 1996) suggested that this particular subfamily of fish transposons might be transferred through even larger evolutionary distances.

Reconstructions of ancestral archetypal genes using parsimony analysis have been reported (Jermann et al., 1995. *Nature* 374, 57-59; Unnikrishnan et al., 1996, Stewart, 1995 *Nature* 374, 12-13). However, such a strategy requires vertical transmission of a gene through evolution for phylogenetically backtracking to the root sequence. Because parsimony analysis could not resolve the phylogenetic relationships between salmonid TcEs, we took the approach of reconstructing a consensus sequence from inactive elements belonging to the same subfamily of transposons. The resurrection of a functional promoter of the L1 retrotransposon in mouse (Adey et al., 1994 *Proc. Natl. Acad. Sci. USA* 91, 1569-1573) has previously been reported.

A strategy for obtaining an active gene is not without risks. The consensus sequence of transposase pseudogenes from a single organism may simply reflect the mutations that had occurred during vertical inactivation that have subsequently been fixed in the genome as a result of amplification of the mutated element. For instance, most Tdr1 elements isolated from zebrafish contain a conserved, 350-bp deletion in the transposase gene (Izsvak et al., 1995). Therefore, their consensus is expected to encode an inactive element. In contrast, because independent fixation of the same mutation in different species is unlikely, we derived a consensus from inactive elements of the same subfamily of transposons from several organisms to provide a sequence for an active transposon.

Both the transposase coding regions and the inverted repeats (IRs) of salmonid-type TcEs accumulated several mutations, including point mutations, deletions and insertions, and show about 5% average pairwise divergence (Ivics et al., 1996, *supra*). Example 1 describes the methods that were used to reconstruct a transposase gene of the salmonid subfamily of fish elements using

the accumulated phylogenetic data. This analysis is provided in the EMBL database as DS30090 from FTP.EBI.AC.AK in directory/pub/databases/embl/align and the product of this analysis was a consensus sequence for an inactive SB protein (SEQ ID NO:2). All the elements that were examined were inactive due to deletions and other mutations. A salmonid transposase gene of the SB transposase family was created using PCR-mutagenesis through the creation of 10 constructs as provided in Fig. 1 and described in Example 1.

This sequence can then be modified further, as described here, to produce active members of the SB protein family.

The SB protein recognizes inverted repeats on a nucleic acid fragment and each inverted repeat includes at least one direct repeat. The gene transfer system of this invention, therefore, comprises two components: a transposase and a cloned, nonautonomous (i.e., non-self inserting) salmonid-type element or transposon (referred to herein as a nucleic acid fragment having at least two inverted repeats) that carries the inverted repeats of the transposon substrate DNA. When put together these two components provide active transposon activity. In use, the transposase binds to the direct repeats in the inverted repeats and promotes integration of the intervening nucleic acid sequence into DNA of a cell a including chromosomes and extra chromosomal DNA of fish as well as mammalian cells. This transposon system does not appear to exist in nature.

The transposase that was reconstructed using the methods of Example 1 represents one member of a family of proteins that can bind to the inverted repeat region of a transposon to effect integration of the intervening nucleic acid sequence into DNA, preferably DNA in a cell. One example of the family of proteins of this invention is provided as SEQ ID NO:1 (see Fig. 2). This family of proteins is referred to herein as SB proteins. The proteins of this invention are provided as a schematic in Fig. 1. The proteins include, from the amino-terminus moving to the carboxy-terminus, a paired-like domain with leucine zipper, one or more nuclear localizing domains (NLS) domains and a catalytic domain including a DD(34)E box and a glycine-rich box as detailed in one example in Fig. 2. The SB family of proteins includes the protein having the amino acid

sequence of SEQ ID NO: 1 and also includes proteins with an amino acid sequence that shares at least an 80% amino acid identity to SEQ ID NO:1. That is, when the proteins of the SB family are aligned, at least 80% of the amino acid sequence is identical. Proteins of the SB family are transposases, that is, they are able to catalyze the integration of nucleic acid into DNA of a cell. In addition, the proteins of this invention are able to bind to the inverted repeat sequences of SEQ ID NOS:4-5 and direct repeat sequences (SEQ ID NOS:6-9) from a transposon as well as a consensus direct repeat sequence (SEQ ID NO:10). The SB proteins preferably have a molecular weight range of about 35 kD to about 40 kD on about a 10% SDS-polyacrylamide gel.

To create an active SB protein, suitable for further modification, a number of chromosomal fragments were sequenced and identified by their homology to the zebrafish transposon-like sequence Tdr1, from eleven species of fish (Ivics et al., 1996). Next these and other homologous sequences were compiled and aligned. The sequences were identified in either GenBank or the EMBL database. Others have suggested using parsimony analysis to arrive at a consensus sequence but in this case parsimony analysis could not resolve the phylogenetic relationships among the salmonid-type TcEs that had been compiled. A consensus transposon was then engineered by changing selected nucleotides in codons to restore the amino acids that were likely to be in that position. This strategy assumes that the most common amino acid in a given position is probably the original (active) amino acid for that locus. The consensus sequence was examined for sites at which it appeared that C->T mutations had been fixed where deamination of ^{5m}C residues may have occurred (which leads to C being converted to T which in turn can lead to the "repair" of the mismatched G residue to an A). In these instances, the "majority-rule" consensus sequence was not always used. Next various expected activities of the resurrected transposase were tested to ensure the accuracy of the engineering.

The amino acid residues described herein employ either the single letter amino acid designator or the three-letter abbreviation. Abbreviations used herein are in keeping with the standard polypeptide nomenclature, *J. Biol. Chem.*, (1969), 243, 3552-3559. All amino acid residue sequences are represented

herein by formulae with left and right orientation in the conventional direction of amino-terminus to carboxy-terminus.

Although particular amino acid sequences encoding the transposases of this invention have been described, there are a variety of conservative changes that can be made to the amino acid sequence of the SB protein without altering SB activity. These changes are termed conservative mutations, that is, an amino acid belonging to a grouping of amino acids having a particular size or characteristic can be substituted for another amino acid, particularly in regions of the protein that are not associated with catalytic activity or DNA binding activity, for example. Other amino acid sequences of the SB protein include amino acid sequences containing conservative changes that do not significantly alter the activity or binding characteristics of the resulting protein. Substitutes for an amino acid sequence may be selected from other members of the class to which the amino acid belongs. For example, the nonpolar (hydrophobic) amino acids include alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, and tyrosine. The polar neutral amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine and glutamine. The positively charged (basic) amino acids include arginine, lysine and histidine. The negatively charged (acidic) amino acids include aspartic acid and glutamic acid. Such alterations are not expected to substantially affect apparent molecular weight as determined by polyacrylamide gel electrophoresis or isoelectric point. Particularly preferred conservative substitutions include, but are not limited to, Lys for Arg and *vice versa* to maintain a positive charge; Glu for Asp and *vice versa* to maintain a negative charge; Ser for Thr so that a free -OH is maintained; and Gln for Asn to maintain a free NH_2.

The SB protein has catalytic activity in a cell but the protein can be introduced into a cell as protein or as nucleic acid. The SB protein can be introduced into the cell as ribonucleic acid, including mRNA; as DNA present in the cell as extrachromosomal DNA including, but not limited to, episomal DNA, as plasmid DNA, or as viral nucleic acid. Further, DNA encoding the SB protein can be stably integrated into the genome of the cell for constitutive or inducible expression. Where the SB protein is introduced into the cell as nucleic acid, the

SB encoding sequence is preferably operably linked to a promoter. There are a variety of promoters that could be used including, but not limited to, constitutive promoters, tissue-specific promoters, inducible promoters, and the like. Promoters are regulatory signals that bind RNA polymerase in a cell to initiate transcription of a downstream (3' direction) coding sequence. A DNA sequence is operably linked to an expression-control sequence, such as a promoter when the expression control sequence controls and regulates the transcription and translation of that DNA sequence. The term "operably linked" includes having an appropriate start signal (e.g., ATG) in front of the DNA sequence to be expressed and maintaining the correct reading frame to permit expression of the DNA sequence under the control of the expression control sequence to yield production of the desired protein product.

One nucleic acid sequence encoding the SB protein is provided as SEQ ID NO:3. In addition to the conservative changes discussed above that would necessarily alter the SB-encoding nucleic acid sequence, there are other DNA or RNA sequences encoding SB protein that have the same amino acid sequence as an SB protein, but which take advantage of the degeneracy of the three letter codons used to specify a particular amino acid. For example, it is well known in the art that the following RNA codons (and therefore, the corresponding DNA codons, with a T substituted for a U) can be used interchangeably to code for each specific amino acid:

Phenylalanine (Phe or F)	UUU or UUC
Leucine (Leu or L)	UUA, UUG, CUU, CUC, CUA or CUG
Isoleucine (Ile or I)	AUU, AUC or AUA
Methionine (Met or M)	AUG
Valine (Val or V)	GUU, GUC, GUA, GUG
Serine (Ser or S)	UCU, UCC, UCA, UCG, AGU, AGC
Proline (Pro or P)	CCU, CCC, CCA, CCG
Threonine (Thr or T)	ACU, ACC, ACA, ACG
Alanine (Ala or A)	GCU, GCG, GCA, GCC
Tyrosine (Tyr or Y)	UAU or UAC
Histidine (His or H)	CAU or CAC

Glutamine (Gln or Q)	CAA or CAG
Asparagine (Asn or N)	AAU or AAC
Lysine (Lys or K)	AAA or AAG
Aspartic Acid (Asp or D)	GAU or GAC
Glutamic Acid (Glu or E)	GAA or GAG
Cysteine (Cys or C)	UGU or UGC
Arginine (Arg or R)	CGU, CGC, CGA, CGG, AGA, AGC
Glycine (Gly or G)	GGU or GGC or GGA or GGG
Termination codon	UAA, UAG or UGA

Further, a particular DNA sequence can be modified to employ the codons preferred for a particular cell type. For example, the preferred codon usage for *E. coli* is known, as are preferred codon usages for animals and humans. These changes are known to those of ordinary skill in the art and are therefore considered part of this invention.

Also contemplated in this invention are antibodies directed to an SB protein of this invention. An "antibody" for purposes of this invention is any immunoglobulin, including antibodies and fragments thereof that specifically binds to an SB protein. The antibodies can be polyclonal, monoclonal and chimeric antibodies. Various methods are known in the art that can be used for the production of polyclonal or monoclonal antibodies to SB protein. See, for example, *Antibodies: A Laboratory Manual*, Harlow and Lane, eds., Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York, 1988).

The nucleic acid encoding the SB protein can be introduced into a cell as a nucleic acid vector such as a plasmid, or as a gene expression vector, including a viral vector. The nucleic acid can be circular or linear. Methods for manipulating DNA and protein are known in the art and are explained in detail in the literature such as Sambrook et al, (1989) *Molecular Cloning: A Laboratory Manual.*, Cold Spring Harbor Laboratory Press or Ausubel, R.M., ed. (1994). *Current Protocols in Molecular Biology*. A vector, as used herein, refers to a plasmid, a viral vector or a cosmid that can incorporate nucleic acid encoding the SB protein or the nucleic acid fragment of this invention. The term "coding sequence" or "open reading frame" refers to a region of nucleic acid that can be

transcribed and/or translated into a polypeptide *in vivo* when placed under the control of the appropriate regulatory sequences.

Another aspect of this invention relates to a nucleic acid fragment, sometimes referred to as a transposon or transposon element that includes a nucleic acid sequence positioned between at least two inverted repeats. Each inverted repeat preferably includes at least one direct repeat (hence, the name IR/DR). The transposon element is a linear nucleic acid fragment (extending from the 5' end to the 3' end, by convention) that can be used as a linear fragment or circularized, for example in a plasmid. In a preferred embodiment there are two direct repeats in each inverted repeat sequence. Preferred direct repeat sequences that bind to SB include:

The 5' outer repeat: (SEQ ID NO:6)

5'-GTTCAAGTCGGAAGTTTACATACACTTAG-3'

The 5' inner repeat: (SEQ ID NO:7)

5'-CAGTGGGTCAGAAGTTTACATACACTAAGG-3'

The 3' inner repeat (SEQ ID NO:8)

5'-CAGTGGGTCAGAAGTTAACATACACTCAATT-3'

The 3' outer repeat (SEQ ID NO:9)

5'-AGTTGAATCGGAAGTTTACATACACCTTAG-3'.

A preferred consensus direct repeat is (SEQ ID NO:10)

5'-CA(GT)TG(AG)GTC(AG)GAAGTTTACATACACTTAAG-3'

In one embodiment the direct repeat sequence includes at least the following sequence:

ACATACAC (SEQ ID NO:11)

A preferred inverted repeat sequence of this invention is SEQ ID NO:4

```
5'-AGTTGAAGTC    GGAAGTTTAC    ATACACTTAA    GTTGGAGTCA
  TTAAAACTCG
   TTTTTCAACT    ACACCACAAA    TTTCTTGTTA    ACAAACAATA
  GTTTTGGCAA
   GTCAGTTAGG    ACATCTACTT    TGTGCATGAC    ACAAGTCATT
  TTTCCAACAA
   TTGTTTACAG    ACAGATTATT    TCACTTATAA    TTCACTGTAT
  CACAATTCCA
   GTGGGTCAGA AGTTTACATA CACTAA-3'
```

and a second inverted repeat sequence of this invention is SEQ ID NO:5

5'-TTGAGTGTAT GTTAACTTCT GACCCACTGG GAATGTGATG AAAGAAATAA
AAGCTGAAAT GAATCATTCT CTCTACTATT ATTCTGATAT TTCACATTCT
TAAAATAAAG TGGTGATCCT AACTGACCTT AAGACAGGGA ATCTTTACTC
GGATTAAATG TCAGGAATTG TGAAAAAGTG AGTTTAAATG TATTTGGCTA
AGGTGTATGT AAACTTCCGA CTTCAACTG-3'.

Preferably the direct repeats are the portion of the inverted repeat that bind to the SB protein to permit insertion and integration of the nucleic acid fragment into the cell. The site of DNA integration for the SB proteins occurs at TA base pairs (see Figure 7B).

The inverted repeats flank a nucleic acid sequence which is inserted into the DNA in a cell. The nucleic acid sequence can include all or part of an open reading from of a gene (i.e., that part of a gene encoding protein), one or more expression control sequences (i.e., regulatory regions in nucleic acid) alone or together with all or part of an open reading frame. Preferred expression control sequences include, but are not limited to promoters, enhancers, border control elements, locus-control regions or silencers. In a preferred embodiment, the nucleic acid sequence comprises a promoter operably linked to at least a portion of an open reading frame.

As illustrated in the examples, the combination of the nucleic acid fragment of this invention comprising a nucleic acid sequence positioned between at least two inverted repeats wherein the inverted repeats can bind to an SB protein and wherein the nucleic acid fragment is capable of integrating into DNA in a cell, in combination with an SB protein (or nucleic acid encoding the SB protein to deliver SB protein to a cell) results in the integration of the nucleic acid sequence into the cell. Alternatively, it is possible for the nucleic acid fragment of this invention to be incorporated into DNA in a cell through non-homologous recombination through a variety of as yet undefined, but reproducible mechanisms. In either event the nucleic acid fragment can be used for gene transfer.

As described in the examples, the SB family of proteins, mediates integration in a variety of cell types and a variety of species. The SB protein

facilitates integration of the nucleic acid fragment of this invention with inverted repeats into both pluripotent (i.e., a cell whose descendants can differentiate into several restricted cell types, such as hematopoietic stem cells or other stem cells) and totipotent cells (i.e., a cell whose descendants can become any cell type in an organism, e.g., embryonic stem cells). It is likely that the gene transfer system of this invention can be used in a variety of cells including animal cells, bacteria, fungi (e.g., yeast) or plants. Animal cells can be vertebrate or invertebrate. Cells such as oocytes, eggs, and one or more cells of an embryo are also considered in this invention. Mature cells from a variety of organs or tissues can receive the nucleic acid fragment of this invention separately, alone, or together with the SB protein or nucleic acid encoding the SB protein. Cells receiving the nucleic acid fragment or the SB protein and capable of receiving the nucleic acid fragment into the DNA of that cell include, but are not limited to, lymphocytes, hepatocytes, neural cells, muscle cells, a variety of blood cells, and a variety of cells of an organism. Example 4 provides methods for determining whether a particular cell is amenable to gene transfer using this invention. The cells can be obtained from vertebrates or invertebrates. Preferred invertebrates include crustaceans or mollusks including, but not limited to shrimp, scallops, lobster, claims, or oysters.

Vertebrate cells also incorporate the nucleic acid fragment of this invention in the presence of the SB protein. Cells from fish, birds and other animals can be used, as can cells from mammals including, but not limited to, rodents, such as rats or mice, ungulates, such as cows or goats, sheep, swine or cells from a human.

The DNA of a cell that acts as a recipient of the nucleic acid fragment of this invention includes any DNA in contact with the nucleic acid fragment of this invention in the presence of an SB protein. For example, the DNA can be part of the cell genome or it can be extrachromosomal, such as an episome, a plasmid, a circular or linear DNA fragment. Targets for integration are double-stranded DNA.

The combination of the nucleic acid fragment of this invention including a nucleic acid sequence positioned between at least two inverted repeats wherein

the inverted repeats can bind to an SB protein and wherein the nucleic acid fragment is capable of integrating into DNA of a cell in combination with a transposase or nucleic acid encoding a transposase, wherein the transposase is an SB protein, including SB proteins that include an amino acid sequence that is 80% identical to SEQ ID NO:1 is useful as a gene transfer system to introduce DNA into the DNA of a cell. In a preferred embodiment, the SB protein comprises the amino acid sequence of SEQ ID NO:1 and in another preferred embodiment the DNA encoding the transposase can hybridize to the DNA of SEQ ID NO:3 under the following hybridization conditions: in 30% (v/v) formamide in 0.5xSSC, 0.1% (w/v) SDS at 42°C for 7 hours.

Gene transfer vectors for gene therapy can be broadly classified as viral vectors or non-viral vectors. The use of the nucleic acid fragment of this invention as a transposon in combination with an SB protein is a refinement of non-viral DNA-mediated gene transfer. Up to the present time, viral vectors have been found to be more efficient at introducing and expressing genes in cells. There are several reasons why non-viral gene transfer is superior to virus-mediated gene transfer for the development of new gene therapies. For example, adapting viruses as agents for gene therapy restricts genetic design to the constraints of that virus genome in terms of size, structure and regulation of expression. Non-viral vectors are generated largely from synthetic starting materials and are therefore more easily manufactured than viral vectors. Non-viral reagents are less likely to be immunogenic than viral agents making repeat administration possible. Non-viral vectors are more stable than viral vectors and therefore better suited for pharmaceutical formulation and application than are viral vectors.

Current non-viral gene transfer systems are not equipped to promote integration of nucleic acid into the DNA of a cell, including host chromosomes. As a result, stable gene transfer frequencies using non-viral systems have been very low; 0.1% at best in tissue culture cells and much less in primary cells and tissues. The present system is a non-viral gene transfer system that facilitates integration and markedly improves the frequency of stable gene transfer.

In the gene transfer system of this invention the SB protein can be

introduced into the cell as a protein or as nucleic acid encoding the protein. In one embodiment the nucleic acid encoding the protein is RNA and in another, the nucleic acid is DNA. Further, nucleic acid encoding the SB protein can be incorporated into a cell through a viral vector, cationic lipid, or other standard transfection mechanisms including electroporation or particle bombardment used for eukaryotic cells. Following introduction of nucleic acid encoding SB, the nucleic acid fragment of this invention can be introduced into the same cell.

Similarly, the nucleic acid fragment can be introduced into the cell as a linear fragment or as a circularized fragment, preferably as a plasmid or as recombinant viral DNA. Preferably the nucleic acid sequence comprises at least a portion of an open reading frame to produce an amino-acid containing product. In a preferred embodiment the nucleic acid sequence encodes at least one protein and includes at least one promoter selected to direct expression of the open reading frame or coding region of the nucleic acid sequence. The protein encoded by the nucleic acid sequence can be any of a variety of recombinant proteins new or known in the art. In one embodiment the protein encoded by the nucleic acid sequence is a marker protein such as green fluorescent protein (GFP), chloramphenicol acetyltransferase (CAT), growth hormones, for example to promote growth in a transgenic animal, β-galactosidase (*lacZ*), luciferase (LUC), and insulin-like growth factors (IGFs).

In one embodiment of a transgenic animal, the protein is a product for isolation from a cell. Transgenic animals as bioreactors are known. Protein can be produced in quantity in milk, urine, blood or eggs. Promoters are known that promote expression in milk, urine, blood or eggs and these include, but are not limited to, casein promoter, the mouse urinary protein promoter, β-globin promoter and the ovalbumin promoter respectively. Recombinant growth hormone, recombinant insulin, and a variety of other recombinant proteins have been produced using other methods for producing protein in a cell. Nucleic acid encoding these or other proteins can be incorporated into the nucleic acid fragment of this invention and introduced into a cell. Efficient incorporation of the nucleic acid fragment into the DNA of a cell occurs when an SB protein is present. Where the cell is part of a tissue or part of a transgenic animal, large

amounts of recombinant protein can be obtained. There are a variety of methods for producing transgenic animals for research or for protein production including, but not limited to (Hackett et al. (1993). The molecular biology of transgenic fish. In *Biochemistry and Molecular Biology of Fishes* (Hochachka & Mommsen, eds) Vol.2, pp. 207-240. Other methods for producing transgenic animals include the teachings of M. Markkula et al., *Rev. Reprod.*, *1*, 97-106 (1996); R. T. Wall et al., *J. Dairy Sci.*, *80*, 2213-2224 (1997); J. C. Dalton, et al., *Adv. Exp. Med. Biol.*, *411*, 419-428 (1997); and H. Lubon et al., *Transfus. Med. Rev.*, *10*, 131-143 (1996). Transgenic zebrafish were made, as described in Example 6. The system has also been tested through the introduction of the nucleic acid with a marker protein into mouse embryonic stem cells (ES) and it is known that these cells can be used to produce transgenic mice (A. Bradley et al., *Nature*, *309*, 255-256 (1984).

In general, there are two methods to achieve improved stocks of commercially important animals. The first is classical breeding, which has worked well for land animals, but it takes decades to make major changes. A review by Hackett et al. (1997) points out that by controlled breeding, growth rates in coho salmon (*Oncorhynchus kisutch*) increased 60% over four generations and body weights of two strains of channel catfish (*Ictalurus punctatus*) were increased 21 to 29% over three generations. The second method is genetic engineering, a selective process by which genes are introduced into the chromosomes of animals or plants to give these organisms a new trait or characteristic, like improved growth or greater resistance to disease. The results of genetic engineering have exceeded those of breeding in some cases. In a single generation, increases in body weight of 58% in common carp (*Cyprinus carpio*) with extra rainbow trout growth hormone I genes, more than 1000% in salmon with extra salmon growth hormone genes, and less in trout were obtained. The advantage of genetic engineering in fish, for example, is that an organism can be altered directly in a very short periods of time if the appropriate gene has been identified (see Hackett, 1997). The disadvantage of genetic engineering in fish is that few of the many genes that are involved in growth and development have been identified and the interactions of their protein products is poorly understood. Procedures for genetic manipulation are lacking

many economically important animals. The present invention provides an efficient system for performing insertional mutagenesis (gene tagging) and efficient procedures for producing transgenic animals. Prior to this invention, transgenic DNA is not efficiently incorporated into chromosomes. Only about one in a million of the foreign DNA molecules integrates into the cellular genome, generally several cleavage cycles into development. Consequently, most transgenic animals are mosaic (Hackett, 1993). As a result, animals raised from embryos into which transgenic DNA has been delivered must be cultured until gametes can be assayed for the presence of integrated foreign DNA. Many transgenic animals fail to express the transgene due to position effects. A simple, reliable procedure that directs early integration of exogenous DNA into the chromosomes of animals at the one-cell stage is needed. The present system helps to fill this need.

The transposon system of this invention has applications to many areas of biotechnology. Development of transposable elements for vectors in animals permits the following: 1) efficient insertion of genetic material into animal chromosomes using the methods given in this application. 2) identification, isolation, and characterization of genes involved with growth and development through the use of transposons as insertional mutagens (e.g., see Kaiser et al., 1995, "Eukaryotic transposable elements as tools to study gene structure and function." In *Mobile Genetic Elements*, IRL Press, pp. 69-100). 3) identification, isolation and characterization of transcriptional regulatory sequences controlling growth and development. 4) use of marker constructs for quantitative trait loci (QTL) analysis. 5) identification of genetic loci of economically important traits, besides those for growth and development, i.e., disease resistance (e.g., Anderson et al., 1996, *Mol. Mar. Biol. Biotech.*, *5*, 105-113). In one example, the system of this invention can be used to produce sterile transgenic fish. Broodstock with inactivated genes could be mated to produce sterile offspring for either biological containment or for maximizing growth rates in aquacultured fish.

In yet another use of the gene transfer system of this invention, the nucleic acid fragment is modified to incorporate a gene to provide a gene therapy to a cell. The gene is placed under the control of a tissue specific promoter or of

a ubiquitous promoter or one or more other expression control regions for the expression of a gene in a cell in need of that gene. A variety of genes are being tested for a variety of gene therapies including, but not limited to, the CFTR gene for cystic fibrosis, adenosine deaminase (ADA) for immune system disorders, factor IX and interleukin-2 (IL-2) for blood cell diseases, alpha-1-antitrypsin for lung disease, and tumor necrosis factors (TNFs) and multiple drug resistance (MDR) proteins for cancer therapies.

These and a variety of human or animal specific gene sequences including gene sequences to encode marker proteins and a variety of recombinant proteins are available in the known gene databases such as GenBank, and the like.

Further, the gene transfer system of this invention can be used as part of a process for working with or for screening a library of recombinant sequences, for example, to assess the function of the sequences or to screen for protein expression, or to assess the effect of a particular protein or a particular expression control region on a particular cell type. In this example, a library of recombinant sequences, such as the product of a combinatorial library or the product of gene shuffling, both techniques now known in the art and not the focus of this invention, can be incorporated into the nucleic acid fragment of this invention to produce a library of nucleic acid fragments with varying nucleic acid sequences positioned between constant inverted repeat sequences. The library is then introduced into cells together with the SB protein as discussed above.

An advantage of this system is that it is not limited to a great extent by the size of the intervening nucleic acid sequence positioned between the inverted repeats. The SB protein has been used to incorporate transposons ranging from 1.3 kilobases (kb) to about 5.0 kb and the *mariner* transposase has mobilized transposons up to about 13 kb. There is no known limit on the size of the nucleic acid sequence that can be incorporated into DNA of a cell using the SB protein.

Rather, what is limiting can be the method by which the gene transfer system of this invention is introduced into cells. For example, where

microinjection is used, there is very little restraint on the size of the intervening sequence of the nucleic acid fragment of this invention. Similarly, lipid-mediated strategies do not have substantial size limitations. However, other strategies for introducing the gene transfer system into a cell, such as viral-mediated strategies could limit the length of the nucleic acid sequence positioned between the inverted repeats, according to this invention.

The two part SB transposon system can be delivered to cells via viruses, including retroviruses (including lentiviruses), adenoviruses, adeno-associated viruses, herpesviruses, and others. There are several potential combinations of delivery mechanisms for the transposon portion containing the transgene of interest flanked by the inverted terminal repeats (IRs) and the gene encoding the transposase. For example, both the transposon and the transposase gene can be contained together on the same recombinant viral genome; a single infection delivers both parts of the SB system such that expression of the transposase then directs cleavage of the transposon from the recombinant viral genome for subsequent integration into a cellular chromosome. In another example, the transposase and the transposon can be delivered separately by a combination of viruses and/or non-viral systems such as lipid-containing reagents. In these cases either the transposon and/or the transposase gene can be delivered by a recombinant virus. In every case, the expressed transposase gene directs liberation of the transposon from its carrier DNA (viral genome) for integration into chromosomal DNA.

This invention also relates to methods for using the gene transfer system of this invention. In one method, the invention relates to the introduction of a nucleic acid fragment comprising a nucleic acid sequence positioned between at least two inverted repeats into a cell. In a preferred embodiment, efficient incorporation of the nucleic acid fragment into the DNA of a cell occurs when the cell also contains an SB protein. As discussed above, the SB protein can be provided to the cell as SB protein or as nucleic acid encoding the SB protein. Nucleic acid encoding the SB protein can take the form of RNA or DNA. The protein can be introduced into the cell alone or in a vector, such as a plasmid or a viral vector. Further, the nucleic acid encoding the SB protein can be stably or

transiently incorporated into the genome of the cell to facilitate temporary or prolonged expression of the SB protein in the cell. Further, promoters or other expression control regions can be operably linked with the nucleic acid encoding the SB protein to regulate expression of the protein in a quantitative or in a tissue-specific manner. As discussed above, the SB protein is a member of a family of SB proteins preferably having at least an 80% amino acid sequence identity to SEQ ID NO:1 and more preferably at least a 90% amino acid sequence identity to SEQ ID NO:1. Further, the SB protein contains a DNA-binding domain, a catalytic domain (having transposase activity) and an NLS signal.

The nucleic acid fragment of this invention is introduced into one or more cells using any of a variety of techniques known in the art such as, but not limited to, microinjection, combining the nucleic acid fragment with lipid vesicles, such as cationic lipid vesicles, particle bombardment, electroporation, DNA condensing reagents (e.g., calcium phosphate, polylysine or polyethyleneimine) or incorporating the nucleic acid fragment into a viral vector and contacting the viral vector with the cell. Where a viral vector is used, the viral vector can include any of a variety of viral vectors known in the art including viral vectors selected from the group consisting of a retroviral vector, an adenovirus vector or an adeno-associated viral vector.

The gene transfer system of this invention can readily be used to produce transgenic animals that carry a particular marker or express a particular protein in one or more cells of the animal. Methods for producing transgenic animals are known in the art and the incorporation of the gene transfer system of this invention into these techniques does not require undue experimentation. The examples provided below teach methods for creating transgenic fish by microinjecting the gene transfer system into a cell of an embryo of the fish. Further, the examples also describe a method for introducing the gene transfer system into mouse embryonic stem cells. Methods for producing transgenic mice from embryonic stem cells are well known in the art. Further a review of the production of biopharmaceutical proteins in the milk of transgenic dairy animals (see Young et al., *BIO PHARM* (1997), 10, 34-38) and the references

provided therein detail methods and strategies for producing recombinant proteins in milk. The methods and the gene transfer system of this invention can be readily incorporated into these transgenic techniques without undue experimentation in view of what is known in the art and particularly in view of this disclosure.

The nucleic acid fragments of this invention in combination with the SB protein or nucleic acid encoding the SB protein is a powerful tool for germline transformation, for the production of transgenic animals, as methods for introducing nucleic acid into DNA in a cell, for insertional mutagenesis, and for gene tagging in a variety of species. Two strategies are diagrammed in Figure 9.

Due to their inherent ability to move from one chromosomal location to another within and between genomes, transposable elements have been exploited as genetic vectors for genetic manipulations in several organisms. Transposon tagging is a technique in which transposons are mobilized to "hop" into genes, thereby inactivating them by insertional mutagenesis. These methods are discussed by Evans et al., *TIG* 1997 13,370-374. In the process, the inactivated genes are "tagged" by the transposable element which then can be used to recover the mutated allele. The ability of the human and other genome projects to acquire gene sequence data has outpaced the ability of scientists to ascribe biological function to the new genes. Therefore, the present invention provides an efficient method for introducing a tag into the genome of a cell. Where the tag is inserted into a location in the cell that disrupts expression of a protein that is associated with a particular phenotype, expression of an altered phenotype in a cell containing the nucleic acid of this invention permits the association of a particular phenotype with a particular gene that has been disrupted by the nucleic acid fragment of this invention. Here the nucleic acid fragment functions as a tag. Primers designed to sequence the genomic DNA flanking the nucleic acid fragment of this invention can be used to obtain sequence information about the disrupted gene.

The nucleic acid fragment can also be used for gene discovery. In one example, the nucleic acid fragment in combination with the SB protein or nucleic acid encoding the SB protein is introduced into a cell. The nucleic acid

fragment preferably comprises a nucleic acid sequence positioned between at least two inverted repeats, wherein the inverted repeats bind to the SB protein and wherein the nucleic acid fragment integrates into the DNA of the cell in the presence of the SB protein. In a preferred embodiment, the nucleic acid sequence includes a marker protein, such as GFP and a restriction endonuclease recognition site, preferably a 6-base recognition sequence. Following integration, the cell DNA is isolated and digested with the restriction endonculease. Where a restriction endonuclease is used that employs a 6-base recognition sequence, the cell DNA is cut into about 4000- bp fragments on average. These fragments can be either cloned or linkers can be added to the ends of the digested fragments to provide complementary sequence for PCR primers. Where linkers are added, PCR reactions are used to amplify fragments using primers from the linkers and primers binding to the direct repeats of the inverted repeats in the nucleic acid fragment. The amplified fragments are then sequenced and the DNA flanking the direct repeats is used to search computer databases such as GenBank.

In another application of this invention, the invention provides a method for mobilizing a nucleic acid sequence in a cell. In this method the nucleic acid fragment of this invention is incorporated into DNA in a cell, as provided in the discussion above. Additional SB protein or nucleic acid encoding the SB protein is introduced into the cell and the protein is able to mobilize (i.e. move) the nucleic acid fragment from a first position within the DNA of the cell to a second position within the DNA of the cell. The DNA of the cell can be genomic DNA or extrachromosomal DNA. The method permits the movement of the nucleic acid fragment from one location in the genome to another location in the genome, or for example, from a plasmid in a cell to the genome of that cell.

All references, patents and publications cited herein are expressly incorporated by reference into this disclosure. Particular embodiments of this invention will be discussed in detail and reference has been made to possible variations within the scope of this invention. There are a variety of alternative

techniques and procedures available to those of skill in the art which would similarly permit one to successfully practice the intended invention.

Example 1
Reconstruction of an SB transposase

Recombinant DNA

Gene reconstruction-Phase 1: Reconstruction of a transposase open reading frame. The Tss1.1 element from Atlantic salmon (GenBank accession number L12206) was PCR-amplified using a primer pair flanking the defective transposase gene, FTC-Start and FTC-Stop to yield product SB1. Next, a segment of the defective transposase gene of the Tss1.2 element (L12207) was PCR-amplified using PCR primers FTC-3 and FTC-4, then further amplified with FTC-3 and FTC-5. The PCR product was digested with restriction enzymes *Nco*I and *Blp*I, underlined in the primer sequences, and cloned to replace the corresponding fragment in SB1 to yield SB2. Then, an approximately 250 bp *Hin*dIII fragment of the defective transposase gene of the Tsg1 element from rainbow trout (L12209) was isolated and cloned into the respective sites in SB2 to result in SB3. The Tss1 and Tsg1 elements were described in (Radice et al., 1994) and were kind gifts from S.W. Emmons.

FTC-Start: 5'-CCTCTAGGATCCGACATCATG (SEQ ID NO:17)
FTC-Stop: 5'-TCTAGAATTCTAGTATTTGGTAGCATTG (SEQ ID NO:18)
FTC-3:5'-AACACCATGGGACCACGCAGCCGTCA (SEQ ID NO:19)
FTC-4:5'-CAGGTTATGTCGATATAGGACTCGTTTTAC (SEQ ID NO:20)
FTC-5:5'-CCTTGCTGAGCGGCCTTTCAGGTTATGTCG (SEQ ID NO:21)

Gene reconstruction-Phase 2: Site-specific PCR mutagenesis of the SB3 open reading frame to introduce consensus amino acids. For PCR mutagenesis, two methods have been used: megaprimer PCR (Sarkar and Sommer, 1990 *BioTechniques* 8, 404-407) from SB4 through SB6, and Ligase Chain Reaction (Michael, 1994 *BioTechniques* 16, 410-412) for steps SB7 to

SB10.

Oligonucleotide primers for product SB4 were the following:

FTC-7: 5'-TTGCACTTTTCGCACCAA for Gln->Arg(74) and Asn->Lys(75) (SEQ ID NO:22);

FTC-13: 5'-GTACCTGTTTCCTCCAGCATC for Ala->Glu(93) (SEQ ID NO:23);

FTC-8: 5'-GAGCAGTGGCTTCTTCCT for Leu->Pro(121) (SEQ ID NO:24);

FTC-9: 5'-CCACAACATGATGCTGCC for Leu->Met(193) (SEQ ID NO:25);

FTC-10: 5'-TGGCCACTCCAATACCTTGAC for Ala->Val(265) and Cys->Trp(268) (SEQ ID NO:26);

FTC-11: 5'-ACACTCTAGACTAGTATTTGGTAGCATTGCC for Ser->Ala(337) and Asn->Lys(339) (SEQ ID NO:27).

Oligonucleotide primers for product SB5:

B5-PTV: 5'-GTGCTTCACGGTTGGGATGGTG for Leu->Pro(183), Asn->Thr(184) and Met->Val(185) (SEQ ID NO:28).

Oligonucleotide primers for product SB6:

FTC-DDE: 5'-ATTTTCTATAGGATTGAGGTCAGGGC for Asp->Glu(279) (SEQ ID NO:29).

Oligonucleotide primers for products SB7 and SB8, in two steps:

PR-GAIS: 5'-GTCTGGTTCATCCTTGGGAGCAATTTCCAAACGCC for Asn->Ile(28), His->Arg(31) and Phe->Ser(21) (SEQ ID NO:30).

Oligonucleotide primers for product SB9:

KARL: 5'-CAAAACCGACATAAGAAAGCCAGACTACGG for Pro->Arg(126) (SEQ ID NO:31);

RA: 5'-ACCATCGTTATGTTTGGAGGAAGAAGGGGGAGGCTTGCAAGCCG for Cys->Arg(166) and Thr->Ala(175) (SEQ ID NO:32);

EY: 5'-GGCATCATGAGGAAGGAAAATTATGTGGATATATTG for Lys->Glu(216) and Asp->Tyr(218) (SEQ ID NO:33);

KRV: 5'-CTGAAAAAGCGTGTGCGAGCAAGGAGGCC for Cys->Arg(288) (SEQ ID NO:34);

VEGYP: 5'-GTGGAAGGCTACCCGAAACGTTTGACC for Leu->Pro(324) (SEQ ID NO:35).

Oligonucleotide primers for product SB10:

FATAH: 5'-GACAAAGATCGTACTTTTTGGAGAAATGTC for Cys->Arg(143) (SEQ ID NO:36).

Plasmids. For pSB10, the SB10 transposase gene was cut with *Eco*RI and *Bam*HI, whose recognition sequences are incorporated and underlined above in the primers FTC-Start and FTC-Stop, filled in with Klenow and cloned into the Klenow-filled *Not*I sites of CMV-βgal (Clonetech), replacing the *LacZ* gene originally present in this plasmid. Because of the blunt-end cloning, both orientations of the gene insert were possible to obtain and the antisense direction was used as a control for transposase. For pSB10-ΔDDE, plasmid pSB10 was cut with *Msc*I, which removes 322 bp of the transposase coding region, and recircularized. Removal of the *Msc*I fragment from the transposase gene deleted much of the catalytic DDE domain and disrupted the reading frame by introducing a premature translational termination codon.

Sequence alignment of 12 partial salmonid-type TcE sequences found in 8 fish species (available under DS30090 from FTP.EBI.AC.AK in directory/pub/databases/embl/align from the EMBL database) allowed us to derive a majority-rule, salmonid-type consensus sequence, and identify conserved protein and DNA sequence motifs that likely have functional importance (Fig. 1A).

Conceptual translation of the mutated transposase open reading frames and comparison with functional motifs in other proteins allowed us to identify five regions that are highly conserved in the SB transposase family (Fig. 1A): I) a paired box/leucine zipper motif at the N-terminus; ii) a DNA-binding domain; iii) a bipartite nuclear localization signal (NLS); iv) a glycine-rich motif close to the center of the transposase without any known function at present; and v) a catalytic domain consisting of three segments in the C-terminal half comprising the DDE domain that catalyzes the transposition. DDE domains were identified by Doak et al. in Tc1 mariner sequences (Doak et al., 1994 *Proc. Natl. Acad. Sci.*

USA 91, 942-946). Multiple sequence alignment also revealed a fairly random distribution of mutations in transposase coding sequences; 72% had occurred at non-synonymous positions in codons. The highest mutation frequencies were observed at CpG dinucleotide sites which are highly mutable (Adey et al., 1994, *supra*). Although amino acid substitutions were distributed throughout the transposases, fewer mutations were detected at the conserved motifs (0.07 non-synonymous mutations per codon), as compared to protein regions between the conserved domains (0.1 non-synonymous mutations per codon). This observation indicated to us that some selection mechanism had maintained the functional domains before inactivation of transposons took place in host genomes. The identification of these putative functional domains was of key importance during the reactivation procedure.

The first step of reactivating the transposase gene, was to restore an open reading frame (SB1 through SB3 in Fig. 1B) from bits and pieces of two inactive TcEs from Atlantic salmon (*Salmo salar*) and a single element from rainbow trout (*Oncorhynchus mykiss*) (Radice et al., 1994, *supra*). SB3, which has a complete open reading frame after removal of stop codons and frameshifts, was tested in an excision assay similar to that described by Handler et al. (1993) but no detectable activity was observed. Due to non-synonymous nucleotide substitutions, the SB3 polypeptide differs from the consensus transposase sequence in 24 positions (Fig. 1B) which can be sorted into two groups; nine residues that are probably essential for transposase activity because they are in the presumed functional domains and/or conserved in the entire Tc1 family, and another fifteen residues whose relative importance could not be predicted. Consequently, we undertook a dual gene reconstruction strategy. First, the putative functional protein domains of the transposase were systematically rebuilt one at a time by correcting the former group of mutations. Each domain for a biochemical activity was tested independently when possible. Second, in parallel with the first approach, a full-length, putative transposase gene was synthesized by extending the reconstruction procedure to all of the 24 mutant amino acids in the putative transposase.

Accordingly, a series of constructs was made to bring the coding

sequence closer, step-by-step, to the consensus using PCR mutagenesis (SB4 through SB10 in Fig. 1B). As a general approach the sequence information predicted by the majority-rule consensus was followed. However, at some codons deamination of ^{5m}C residues of CpG sites occurred, and C ->T mutations had been fixed in many elements. At R(288), where TpG's and CpG's were represented in equal numbers in the alignment, the CpG sequence was chosen because the CpG ->TpG transition is more common in vertebrates than the TpG -> CpG. The result of this extensive genetic engineering is a synthetic transposase gene encoding 340 amino acids (SB10 in Figs. 1B and 2).

The reconstituted functional transposase domains were tested for activity. First, a short segment of the SB4 transposase gene (Fig. 1B) encoding an NLS-like protein motif was fused to the *lacZ* gene. The transposase NLS was able to mediate the transfer of the cytoplasmic marker-protein, β-galactosidase, into the nuclei of cultured mouse cells (Ivics et al., 1996, *supra*), supporting our predictions that a bipartite NLS was a functional motif in SB and that our approach to resurrect a full-length, multifunctional enzyme was viable.

Example 2

Preparation of a nucleic acid fragment with inverted repeat sequences.

In contrast to the prototypic Tc1 transposon from *Caenorhabditis elegans* which has short, 54-bp indirect repeat sequences (IRs) flanking its transposase gene, most TcEs in fish belong to the IR/DR subgroup of TcEs (Ivics et al., 1996; Izsvak et al., 1995, both *supra*) which have long, 210-250 bp IRs at their termini and directly repeated DNA sequence motifs (DRs) at the ends of each IR (Fig. 1A). However, the consensus IR sequences are not perfect repeats (i.e., similar, but not identical) indicating that, in contrast to most TcEs, these fish elements naturally possess imperfect inverted repeats. The match is less than 80% at the center of the IRs, but is perfect at the DRs, suggesting that this nonrandom distribution of dissimilarity could be the result of positive selection that has maintained functionally important sequence motifs in the IRs (Fig. 3). Therefore, we suspected that DNA sequences at and around the DRs might carry *cis*-acting information for transposition and mutations within the IRs, but outside

the DRs, would probably not impair the ability of the element to transpose. As a model substrate, we chose a single salmonid-type TcE substrate sequence from *Tanichthys albonubes* (hereafter referred to as *T*) which has intact DR motifs whose sequences are only 3.8% divergent from the salmonid consensus. The variation in the DNase-protected regions of the four DR sequences varied from about 83% to about 95 %, see SEQ ID NOS:6-9.

A TcE from *Tanichthys albonubes* (L48685) was cloned into the *Sma*I site of pUC19 to result in pT. The donor construct for the integration assays, pT/neo, was made by cloning, after Klenow fill-in, an *Eco*RI/*Bam*HI fragment of the plasmid pRc-CMV (Invitrogen, San Diego, CA) containing the SV40 promoter/enhancer, the neomycin resistance gene and an SV40 poly(A) signal into the *Stu*I/*Msc*I sites of pT. The *Stu*I/*Msc*I double digest of *T* leaves 352 bp on the left side and 372 bp on the right side of the transposon and thus contains the terminal inverted repeats. An *Eco*RI digest of pT/neo removed a 350 bp fragment including the left inverted repeat of the transposon, and this plasmid, designated pT/neo-ΔIR, was used as a control for the substrate-dependence of transposase-mediated transgene integration (see Example 4)

Example 3

DNA specificity of an SB transposase

There are at least two distinct subfamilies of TcEs in the genomes of Atlantic salmon and zebrafish, Tss1/Tdr1 and Tss2/Tdr2, respectively. Elements from the same subfamily are more alike, having about 70% nucleic acid identity, even when they are from two different species (e.g., Tss1 and Tdr1) than members of two different subfamilies in the same species. For example, Tdr1 and Tdr2 are characteristically different in their encoded transposases and their inverted repeat sequences, and share only about 30% nucleic acid identity. It may be that certain subfamilies of transposons must be significantly different from each other in order to avoid cross-mobilization. A major question is whether substrate recognition of transposases is sufficiently specific to prevent activation of transposons of closely related subfamilies.

We have shown that the 12-bp DRs of salmonid-type elements, identical

to the DRs of zebrafish-type TcEs, are part of the binding sites for SB. However, these binding-sites are 30 bp long. Thus, specific DNA-binding also involves DNA sequences around the DRs that are variable between TcE subfamilies in fish. Such a difference in the sequences of transposase binding sites might explain the inability of N123 to bind efficiently to zebrafish Tdr1 IRs, and may enable the transposase to distinguish even between closely related TcE subfamilies. Indeed, mutations of four base pairs in the 20-bp Tc1 binding site can abolish binding of transposase (Vos and Plasterk, 1994 *EMBO J.* 13, 6125-6132). The DR core motifs are likely involved primarily in transposase-binding while sequences around the DR motifs likely provide the specificity for this binding.

SB has four binding-sites in its transposon substrate DNA that are located at the ends of the IRs. These sites share about a 83% to about a 95% identity (by comparison of SEQ ID NOS:6-9). However, a zebrafish Tdr1 element lacking an internal transposase-binding site was apparently able to transpose. This observation agrees with the finding that removal of internal transposase-binding sites from engineered Tc3 elements did not lessen their ability to transpose (Colloms et al., 1994 *Nucl. Acids Res.* 22, 5548-5554), suggesting that the presence of internal transposase-binding sites is not essential for transposition. Multiple binding-sites for proteins, including transposases, are frequently associated with regulatory functions (Gierl et al., 1988 *EMBO J.* 7, 4045-4053). Consequently, the internal binding-sites for transposases in the IR/DR group of TcEs serve one or more regulatory purposes affecting transposition and/or gene expression.

Once in the nucleus, a transposase must bind specifically to its recognition sequences in the transposon. The specific DNA-binding domains of both the Tc1 and Tc3 transposases have been mapped to their N-terminal regions (Colloms et al., 1994, *supra*; Vos and Plasterk, 1994, *supra*). However, there is very little sequence conservation between the N-terminal regions of TcE transposases, suggesting that these sequences are likely to encode specific DNA-binding functions in these proteins. On the other hand, the N-terminal region of SB has significant structural and sequence similarities to the *paired* DNA-

binding domain, found in the Pax family of transcription factors, in a novel combination with a leucine zipper-like motif (Ivics et al., 1996, *supra*). A gene segment encoding the first 123 amino acids of SB (N123), which presumably contains all the necessary information for specific DNA-binding and includes the NLS, was reconstructed (SB8 in Fig. 1B), and expressed in *E. coli*. N123 was purified via a C-terminal histidine tag as a 16 KDa polypeptide (Fig. 3A).

Induction of N123 was in *E. coli* strain BL21(DE3) (Novagen) by the addition of 0.4 mM IPTG at 0.5 O.D. at 600 nm and continued for 2.5 h at 30°C. Cells were sonicated in 25 mM HEPES, pH 7.5, 1 M NaCl, 15% glycerol, 0.25% Tween 20, 2 mM β-mercaptoethanol, 1 mM PMSF) and 10 mM imidazole (pH 8.0) was added to the soluble fraction before it was mixed with Ni^{2+}-NTA resin (Qiagen) according to the recommendations of the manufacturer. The resin was washed with 25 mM HEPES (pH 7.5), 1 M NaCl, 30% glycerol, 0.25% Tween 20, 2 mM β–mercaptoethanol, 1 mM PMSF and 50 mM imidazole (pH 8.0) and bound proteins were eluted with sonication buffer containing 300 mM imidazole, and dialyzed overnight at 4°C against sonication buffer without imidazole.

In addition to the NLS function, N123 also contains the specific DNA-binding domain of SB, as tested in a mobility-shift assay (Fig. 3B). A 300 bp *Eco*RI/*Hin*dIII fragment of pT comprising the left inverted repeat of the element was end-labeled using [α^{32}P]dCTP and Klenow. Nucleoprotein complexes were formed in 20 mM HEPES (pH 7.5), 0.1 mM EDTA, 0.1 mg/ml BSA, 150 mM NaCl, 1 mM DTT in a total volume of 10 μl. Reactions contained 100 pg labeled probe, 2 μg poly[dI][dC] and 1.5 μl N123. After 15 min incubation on ice, 5 μl of loading dye containing 50% glycerol and bromophenol blue was added and the samples loaded onto a 5% polyacrylamide gel (Ausubel). DNaseI footprinting was done using a kit from BRL according to the recommendations of the manufacturer. Upon incubation of a radiolabeled 300-bp DNA fragment comprising the left IR of *T*, deoxyribonucleoprotein complexes were observed (Fig. 3B, left panel- lane 3), as compared to samples containing extracts of bacteria transformed with the expression vector only (lane 2) or probe without any protein (lane 1). Unlabelled IR sequences of *T*, added in excess to the reaction as competitor DNA, inhibited binding of the probe (lane 4), whereas the

analogous region of a cloned Tdr1 element from zebrafish did not appreciably compete with binding (lane 5). Thus, N123 is able to distinguish between salmonid-type and zebrafish-type TcE substrates.

The number of the deoxyribonucleoprotein complexes detected by the mobility-shift assay at increasingly higher N123 concentrations indicated two protein molecules bound per IR (Fig. 3B, right panel), consistent with either two binding sites for transposase within the IR or a transposase dimer bound to a single site. Transposase-binding sites were further analyzed and mapped in a DNaseI footprinting experiment. Using the same fragment of *T* as above, two protected regions close to the ends of the IR probe were observed (Fig. 4). The two 30-bp footprints cover the subterminal DR motifs within the IRs. Thus, the DRs are the core sequences for DNA-binding by N123. The DR motifs are almost identical between salmonid- and zebrafish-type TcEs (Ivics et al., 1997). However, the 30-bp transposase binding-sites are longer than the DR motifs and contain 8 base pairs and 7 base pairs in the outer and internal binding sites, respectively, that are different between the zebrafish- and the salmonid-type IRs (Fig. 4B).

Although there are two binding-sites for transposase near the ends of each IR, apparently only the outer sites are utilized for DNA cleavage and thus excision of the transposon. Sequence comparison shows that there is a 3-bp difference in composition and a 2-bp difference in length between the outer and internal transposase-binding sites (Fig. 4C). In summary, our synthetic transposase protein has DNA-binding activity and this binding appears to be specific for salmonid-type IR/DR sequences.

For the expression of an N-terminal derivative of SB transposase, a gene segment of SB8 was PCR-amplified using primers FTC-Start and FTC-8, 5'-phosphorylated with T4 polynucleotide kinase, digested with *Bam*HI, filled in with Klenow, and cloned into the *NdeI*/*Eco*RI digested expression vector pET21a (Novagen) after Klenow fill-in. This plasmid, pET21a/N123 expresses the first 123 amino acids of the transposase (N123) with a C-terminal histidine tag.

Example 4

Transposition of DNA by an SB transposase

The following experiments demonstrate that the synthetic, salmonid-type SB transposase performed all of the complex steps of transposition, i.e., recognized a DNA molecule, excised the substrate DNA and inserted it into the DNA of a cell, such as a cell chromosome. This is in contrast to control samples that did not include the SB transposase and therefore measured integration through non-homologous recombination.

Upon cotransfection of the two-component SB transposon system into cultured vertebrate cells, transposase activity manifested as enhanced integration of the transgene serving as the DNA substrate for transposase. The binding of transposase to a donor construct and subsequent active transport of these nucleoprotein complexes into the nuclei of transfected cells could have resulted in elevated integration rates, as observed for transgenic zebrafish embryos using an SV40 NLS peptide (Collas et al., 1996 *Transgenic Res.* 5, 451-458). However, DNA-binding and nuclear targeting activities alone did not increase transformation frequency, which occurred only in the presence of full-length transposase. Although not sufficient, these functions are probably necessary for transposase activity. Indeed, a single amino acid replacement in the NLS of *mariner* is detrimental to overall transposase function (Lohe et al., 1997 *Proc. Natl. Acad. Sci. USA* 94, 1293-1297). The inability of SB6, a mutated version of the transposase gene, to catalyze transposition demonstrates the importance of the sequences of the conserved motifs. Notably, three of the 11 amino acid substitutions that SB6 contains, F(21), N(28) and H(31) are within the specific DNA-binding domain (Figs. 1 and 2). Sequence analysis of the *paired*-like DNA-binding domain of fish TcE transposases indicates that an isoleucine at position 28 is conserved between the transposases and the corresponding positions in the Pax proteins (Ivics et al., 1996, *supra*). Thus, we predict that this motif is crucial for DNA-binding activity. SB exhibits substrate-dependence for specific recognition and integration; only those engineered transposons that have both of the terminal inverted repeats can be transposed by SB. Similarly, in P element transformation in *Drosophila*, the transposase-producing helper

construct is often a "wings-clipped" transposase gene which lacks one of the inverted repeats of P which prevents the element from jumping (Cooley et al., 1988 *Science* 239, 1121-1128). In our transient assay, transposition can only occur if both components of the SB system are present in the same cell. Once that happens, multiple integrations can take place as demonstrated by our finding of up to 11 integrated transgenes in neomycin-resistant cell clones (Fig. 7A). In contrast to spontaneous integration of plasmid DNA in cultured mammalian cells that often occurs in the form of concatemeric multimers into a single genomic site (Perucho et al., 1980 *Cell* 22, 309-317), these multiple insertions appear to have occurred in distinct chromosomal locations.

Integration of our synthetic, salmonid transposons was observed in fish as well as in mouse and human cells. In addition, recombination of genetic markers in a plasmid-to-plasmid transposition assay (Lampe et al., 1996, *supra*) was significantly enhanced in microinjected zebrafish embryos in the presence of transposase. Consequently, SB apparently does not need any obvious, species-specific factor that would restrict its activity to its original host. Importantly, the most significant enhancement, about 20-fold, of transgene integration was observed in human cells as well as fish embryonic cells.

Integration activity of *SB*

In addition to the abilities to enter nuclei and specifically bind to its sites of action within the inverted repeats, a fully active transposase is expected to excise and integrate transposons. In the C-terminal half of the SB transposase, three protein motifs make up the DD(34)E catalytic domain; the two invariable aspartic acid residues, D(153) and D(244), and a glutamic acid residue, E(279), the latter two being separated by 34 amino acids (Fig. 2). An intact DD(34)E box is essential for catalytic functions of Tc1 and Tc3 transposases (van Luenen et al., 1994 *Cell* 79, 293-301; Vos and Plasterk, 1994, *supra*).

Two different integration assays were used. A first assay was designed to detect chromosomal integration events into the chromosomes of cultured cells. The assay is based on *trans*-complementation of two nonautonomous transposable elements, one containing a selectable marker gene (donor) and another that expresses the transposase (helper) (Fig. 5A). The donor, pT/neo, is

an engineered, *T*-based element which contains an SV40 promoter-driven *neo* gene flanked by the terminal IRs of the transposon containing binding sites for the transposase. The helper construct expresses the full-length SB10 transposase gene driven by a human cytomegalovirus (CMV) enhancer/promoter. In the assay, the donor plasmid is cotransfected with the helper or control constructs into cultured vertebrate cells, and the number of cell clones that are resistant to the neomycin analog drug G-418 due to chromosomal integration and expression of the neo transgene serves as an indicator of the efficiency of gene transfer. If SB is not strictly host-specific, transposition should also occur in phylogenetically distant vertebrate species. Using the assay system shown in Fig. 5A, enhanced levels of transgene integration were observed in the presence of the helper plasmid; more than 5-fold in mouse LMTK cells and more than 20-fold in human HeLa cells (Figs. 5B and 6). Consequently, SB appears to be able to increase the efficiency of transgene integration, and this activity is not restricted to fish cells.

To analyze the requirements for enhanced transgene integration, further experiments were conducted. Fig. 5B shows five plates of transfected HeLa cells that were placed under G-418 selection, and were stained with methylene blue two weeks post-transfection. The staining patterns clearly demonstrate a significant increase in integration of neo-marked transposons into the chromosomes of HeLa cells when the SB transposase-expressing helper construct was cotransfected (plate 2), as compared to a control cotransfection of the donor plasmid plus the SB transposase gene cloned in an antisense orientation (pSB10-AS; plate 1). This result indicates that the production of transposase protein was essential for enhanced chromosomal integration of the transgene and demonstrates that the transposase is precise even in human cells.

In a second assay, an indicator plasmid containing the transposase recognition sequence and a marker gene (Ampicillin resistance) was co-injected with a target plasmid containing a kanamycin gene and SB transposase. Resulting plasmids were isolated and used to transform *E. coli*. Colonies were selected for ampicillin and kanamycin resistance (see Figure 8). While SB transposase was co-microinjected in these assays, mRNA encoding the SB

transposase could also be co-microinjected in place of or in addition to, the SB transposase protein.

Cell transfections

Cells were cultured in DMEM supplemented with 10% fetal bovine serum, seeded onto 6 cm plates one day prior to transfection and transfected with 5 μg Elutip (Schleicher and Schuell)-purified plasmid DNA using Lipofectin from BRL. After 5 hrs of incubation with the DNA-lipid complexes, the cells were "glycerol-shocked" for 30 sec with 15% glycerol in phosphate buffered saline (PBS), washed once with PBS and then refed with serum-containing medium. Two days post-transfection, the transfected cells were trypsinized, resuspended in 2 ml of serum-containing DMEM and either 1 ml or 0.1 ml aliquots of this cell suspension were seeded onto several 10 cm plates in medium containing 600 μg/ml G-418 (BRL). After two weeks of selection, cell clones were either picked and expanded into individual cultures or fixed with 10% formaldehyde in PBS for 15 min, stained with methylene blue in PBS for 30 min, washed extensively with deionized water, air dried and photographed.

These assays can also be used to map transposase domains necessary for chromosomal integration. For this assay, a frameshift mutation was introduced into the SB transposase gene which put a translational stop codon behind G(161). This construct, pSB10-ΔDDE, expresses a truncated transposase polypeptide that contains specific DNA-binding and NLS domains, but lacks the catalytic domain. The transformation rates obtained using this construct (plate 3 in Fig. 5B) were similar to those obtained with the antisense control (Fig. 6). This result suggests that the presence of a full-length transposase protein is necessary and that DNA-binding and nuclear transport activities themselves are not sufficient for the observed enhancement of transgene integration.

As a further control of transposase requirement, the integration activity of an earlier version of the SB transposase gene was tested, SB6 which differs from SB10 at 11 residues, Fig. 1B), using the same assay. The number of transformants observed using SB6 (plate 4 in Fig. 5B) was about the same as with the antisense control experiment (Fig. 6), indicating that the amino acid replacements that we introduced into the transposase gene were critical for

transposase function. In summary, the three controls shown in plates 1, 3, and 4 of Fig. 5B establish the *trans*-requirements of enhanced, SB-mediated transgene integration.

True transposition requires a transposon with intact IR sequences. One of the IRs of the *neo*-marked transposon substrate was removed, and the performance of this construct, pT/neo-ΔIR, was tested for integration. The transformation rates observed with this plasmid (plate 5 in Fig. 5B) were more than 7-fold lower than those with the full-length donor (Fig. 6). These results indicated that both of the IRs flanking the transposon are required for efficient transposition and thereby establish some of the *cis*-requirements of the two-component SB transposon system.

To examine the structures of integrated transgenes, eleven colonies of cells growing under G-418 selection from an experiment similar to that shown in plate 2 in Fig. 5B were picked and their DNAs analyzed using Southern hybridization. Genomic DNA samples of the cell clones were digested with a combination of five restriction enzymes that do not cut within the 2233 bp T/neo marker transposon, and hybridized with a *neo*-specific probe (Fig. 7). The hybridization patterns indicated that all of the analyzed clones contained integrated transgenes in the range of 1 (lane 4) to 11 (lane 2) copies per transformant. Moreover, many of the multiple insertions appear to have occurred in different locations in the human genome.

The presence of duplicated TA sequences flanking an integrated transposon is a hallmark of TcE transposition. To reveal such sequences, junction fragments of integrated transposons and human genomic DNA were isolated using a ligation-mediated PCR assay (Devon et al., *Nucl. Acids. Res.*, *23*, 1644-1645 (1995)). We have cloned and sequenced junction fragments of five integrated transposons, all of them showing the predicted sequences of the IRs which continue with TA dinucleotides and sequences that are different in all of the junctions and different from the plasmid vector sequences originally flanking the transposon in pT/neo (Fig. 7B). The same results were obtained from nine additional junctions containing either the left or the right IR of the transposon (data not shown). These results indicated that the marker transposons had been precisely excised from the

donor plasmids and subsequently spliced into various locations in human chromosomes. Next, the junction sequences were compared to the corresponding "empty" chromosomal regions cloned from wild-type HeLa DNA. As shown in Fig. 7B, all of these insertions had occurred into TA target sites, which were subsequently duplicated to result in TA's flanking the integrated transposons. These data demonstrate that SB uses the same, cut-and-paste-type mechanism of transposition as other members of the Tc1/*mariner* superfamily and that fidelity of the reaction is maintained in heterologous cells. These data also suggest that the frequency of SB-mediated transposition is at least 15-fold higher than random recombination. Since none of the sequenced recombination events were mediated by SB-transposase, the real rate of transposition over random recombination could be many fold higher. If the integration is the result of random integration that was not mediated by the SB protein, the ends of the inserted neo construct would not correspond to the ends of the plasmids; there would have been either missing IR sequences and/or additional plasmid sequences that flank the transposon. Moreover, there would not have been duplicated TA base-pairs at the sites of integration.

Taken together, the dependence of excision and integration, from extrachromosomal plasmids to the chromosomes of vertebrate cells, of a complete transposon with inverted repeats at both ends by a full-length transposase enzyme demonstrates that the gene transfer system is fully functional.

Example 5

Transposition of DNA in cells from different species

Host-requirements of transposase activity were assessed using three different vertebrate cells, LMTK from mouse and HeLa from human and embryonic cells from the zebarafish.

An assay was designed to demonstrate that the transposase worked in a functioning set of cells (i.e., embryonic cells that were differentiating and growing in a natural environment). The assay involved inter-plasmid transfer where the transposon in one plasmid is removed and inserted into a target

plasmid and the transposase construct was injected into 1-cell stage zebra fish embryos. In these experiments the Indicator (donor) plasmids for monitoring transposon excision and/or integration included: 1) a marker gene that when recovered in *E. coli* or in fish cells, could be screened by virtue of either the loss or the gain of a function, and 2) transposase-recognition sequences in the IRs flanking the marker gene. The total size of the marked transposons was kept to about 1.6 kb, the natural size of the TcEs found in teleost genomes. However, the rate of gene transfer using transposons of about 5 kb is not significantly different from that for the 1.6 kb transposon, suggesting that transposition can occur with large transposons. The transposition activity of Ts1 transposase was evaluated by co-microinjecting 200 ng/µl of Ts1 mRNA, made in vitro by T7 RNA polymerase from a Bluescript expression vector, plus about 250 ng/µl each of target and donor plasmids into 1-cell stage zebrafish embryos. Low molecular weight DNA was prepared from the embryos at about 5 hrs post-injection, transformed into *E.coli* cells, and colonies selected by replica plating on agar containing 50 µg/ml kanamycin and/or ampicillin. In these studies there was a transposition frequency into the target plasmid was about 0.041% in experimental cells as compared to 0.002% in control cells. This level did not include transpositions that occurred in the zebrafish genome. In these experiments we found that about 40% to 50% of the embryos did not survive beyond 4 days. Insertional mutagenesis studies in the mouse have suggested that the rate of recessive lethality is about 0.05 (i.e., an average of about 20 insertions will be lethal). Assuming that this rate is applicable to zebrafish, the approximate level of mortality suggests that with the microinjection conditions used in these experiments, about 20 insertions per genome, the mortality can be accounted for.

Example 6

Stable gene expression from SB transposons

A transposon system will be functional for gene transfer, for such purposes as gene therapy and gene delivery to animal chromosomes for bioreactor systems,

only if the delivered genes are reliably expressed. To determine the fidelity of gene expression following *Sleeping Beauty* transposase-mediated delivery, we co-microinjected a transposon containing the Green Fluorescent Protein (GFP) gene under the direction of an SV40 promoter plus in vitro-synthesized mRNA encoding *Sleeping Beauty* transposase into 1-cell zebrafish embryos. 34 of the injected embryos, that showed some expression of GFP during embryogenesis, were allowed to grow to maturity and were mated with wild-type zebrafish. From these matings we found that 4 of the 34 fish could transfer a GFP gene to their progeny (Table 1). The expression of GFP in the offspring of these four F0 fish, identified as A, B, C, and D, was evaluated and the fish were grown up. From the original four founders, the rate of transmission of the GFP gene ranged from about 2% to 12% (Table 1), with an average of about 7%. The expression of GFP in these fish was nearly the same in all individuals in the same tissue types, suggesting that expression of the GFP gene could be revived following transmission through eggs and sperm. These data suggest that the germ-lines were mosaic for expressing GFP genes and that the expression of the genes was stable. The F1 offspring of Fish D were mated with each other. In this case we would expect about 75% transmission and we found that indeed 69/90 (77%) F2 fish expressed the GFP protein at comparable levels in the same tissues; further testimony of the ability of the SB transposon system to deliver genes that can be reliably expressed through at least two generations of animal.

Table 1
Stability of gene expression in zebrafish following injection of a SB transposon containing the GFP gene.

Transgenic Line	Expression of GFP F0	F1	F2
34 founders	34 (of which 4 progeny, A-D, passed on the transgene)		
A		25/200 (12%)	
B		76/863 (9%)	
C		12/701 (2%)	
D		86/946 (10%)	69/90 (77%)

The numbers in the columns for fish A-D show the numbers of GFP expressing fish followed by the total number of offspring examined. The percentages of GFP-expressing offspring are given in parentheses.

Example 7

SB Transposons for Insertional Mutagenesis and Gene Discovery

Due to their inherent ability to move from one chromosomal location to another within and between genomes, transposable elements have revolutionized genetic manipulation of certain organisms including bacteria (Gonzales et al., 1996 *Vet. Microbiol.* 48, 283-291; Lee and Henk, 1996. *Vet. Microbiol.* 50, 143-148), *Drosophila* (Ballinger and Benzer, 1989 *Proc. Natl. Acad. Sci. USA* 86, 9402-9406; Bellen et al., 1989 *Genes Dev.* 3, 1288-1300; Spradling et al., 1995 *Proc. Natl. Acad. Sci. USA* 92, 10824-10830), *C. elegans* (Plasterk, 1995. *Meth. Cell. Biol.*, Academic Press, Inc. pp. 59-80) and a variety of plant species (Osborne and Baker, *Curr. Opin. Cell Biol*, 7, 406-413 (1995)). Transposons have been harnessed as useful vectors for transposon-tagging, enhancer trapping and transgenesis. However, the majority, if not all, animals of economic importance lack such a tool. For its simplicity and apparent ability to function in diverse organisms, *SB* should prove useful as an efficient vector for species in which DNA transposon technology is currently not available.

An SB-type transposable element can integrate into either of two types of chromatin, functional DNA sequences where it may have a deleterious effect due to insertional mutagenesis or non-functional chromatin where it may not have much of a consequence (Fig. 9). This power of "transposon tagging" has been exploited in simpler model systems for nearly two decades (Bingham *et al.*, *Cell*, *25*, 693-704 (1981); Bellen *et al.*, 1989, *supra*). Transposon tagging is an old technique in which transgenic DNA is delivered to cells so that it will integrate into genes, thereby inactivating them by insertional mutagenesis. In the process, the inactivated genes are tagged by the transposable element which then can be used to recover the mutated allele. Insertion of a transposable element may disrupt the function of a gene which can lead to a characteristic phenotype. As illustrated in Fig. 10, because insertion is approximately random, the same procedures that generate insertional, loss-of-function mutants can often be used to deliver genes that will confer new phenotypes to cells. Gain-of-function mutants can be used to understand the roles that gene products play in growth and development as well as the importance of their regulation.

There are several ways of isolating the tagged gene. In all cases genomic DNA is isolated from cells from one or more tissues of the mutated animal by conventional techniques (which vary for different tissues and animals). The DNA is cleaved by a restriction endonuclease that may or may not cut in the transposon tag (more often than not it does cleave at a known site). The resulting fragments can then either be directly cloned into plasmids or phage vectors for identification using probes to the transposon DNA (see Kim et al., 1995 for references in *Mobile Genetic Elements*, IRL Press, D. L. Sheratt eds.). Alternatively, the DNA can be PCR amplified in any of many ways; we have used the LM-PCR procedure of Izsvak and Ivics (1993, *supra*) and a modification by Devon et al. (1995, *supra*) and identified by its hybridization to the transposon probe. An alternative method is inverse-PCR (e.g., Allende et al., *Genes Dev.*, *10*, 3141-3155 (1996)). Regardless of method for cloning, the identified clone is then sequenced. The sequences that flank the transposon (or other inserted DNA) can be identified by their non-identity to the insertional element. The sequences can be combined and then used to search the nucleic acid databases for either homology with other

previously characterized gene(s), or partial homology to a gene or sequence motif that encodes some function. In some cases the gene has no homology to any known protein. It becomes a new sequence to which others will be compared. The encoded protein will be the center of further investigation of its role in causing the phenotype that induced its recovery.

Example 8

SB transposons as markers for gene mapping

Repetitive elements for mapping transgenes and other genetic loci have also been identified. DANA is a retroposon with an unusual substructure of distinct cassettes that appears to have been assembled by insertions of short sequences into a progenitor SINE element. DANA has been amplified in the *Danio* lineage to about 4 x 10^5 copies/genome. *Angel* elements, which are nearly as abundant as DANA, are inverted-repeat sequences that are found in the vicinity of fish genes. Both DANA and *Angel* elements appear to be randomly distributed in the genome, and segregate in a Medelian fashion. PCR amplifications using primers specific to DANA and *Angel* elements can be used as genetic markers for screening polymorphisms between fish stocks and localization of transgenic sequences. Interspersed repetitive sequence-PCR (IRS-PCR) can be used to detect polymorphic DNA. IRS-PCR amplifies genomic DNA flanked by repetitive elements, using repeat-specific primers to produce polymorphic fragments that are inherited in a Medelian fashion (Fig. 10A). Polymorphic DNA fragments can be generated by DANA or Angel specific primers in IRS-PCR and the number of detectable polymorphic bands can be significantly increased by the combination of various primers to repetitive sequences in the zebrafish genome, including SB-like transposons.

Polymorphic fragments can be recovered from gels and cloned to provide sequence tagged sites (STSs) for mapping mutations. Fig. 10B illustrates the general principles and constraints for using IRS-PCR to generate STSs. We estimate that about 0.1% of the zebrafish genome can be directly analyzed by IRS-PCR using only 4 primers. The four conserved (C1-4) regions of DANA seem to have different degrees of conservation and representation in the

zebrafish genome and this is taken into account when designing PCR primers.

The same method has a potential application in fingerprinting fish stocks and other animal populations. The method can facilitate obtaining subclones of large DNAs cloned in yeast, bacterial and bacteriophage P1-derived artificial chromosomes (YACs, BACs and PACs respectively) and can be used for the detection of integrated transgenic sequences.

It will be appreciated by those skilled in the art that while the invention has been described above in connection with particular embodiments and examples, the invention is not necessarily so limited and that numerous other embodiments, examples, uses, modifications and departures from the embodiments, examples and uses may be made without departing from the inventive scope of this application.

What is Claimed is:

1. A nucleic acid fragment comprising:

a nucleic acid sequence positioned between at least two inverted repeats wherein the inverted repeats can bind to an SB protein and wherein the nucleic acid fragment is capable of integrating into DNA in a cell.

2. The fragment of claim 1 wherein the nucleic acid fragment is part of a plasmid.

3. The fragment of claim 1 wherein the nucleic acid sequence comprises at least a portion of an open reading frame.

4. The fragment of claim 1 wherein the nucleic acid sequence comprises at least one expression control region.

5. The fragment of claim 4 wherein the expression control region is selected from the group consisting of a promoter, an enhancer or a silencer.

6. The fragment of claim 1 wherein the nucleic acid sequence comprises a promoter operably linked to at least a portion of an open reading frame.

7. The fragment of claim 1 wherein the cell is obtained from an animal

8. The fragment of claim 7 wherein the cell is obtained from an invertebrate.

9. The fragment of claim 8 wherein the invertebrate is a crustacean or a mollusk.

10. The fragment of claim 9 wherein the crustacean or mollusk is a shrimp, a scallop, a lobster or an oyster.

11. The fragment of claim 7 wherein the cell is obtained from a vertebrate.

12. The fragment of claim 10 wherein the cell is obtained from a fish.

13. The fragment of claim 11 wherein the cell is obtained from a bird.

14. The fragment of claim 11 wherein the vertebrate is a mammal.

15. The fragment of claim 14 wherein the cell is obtained from the group consisting of mice, ungulates, sheep, swine, and humans.

16. The fragment of claim 1 wherein the DNA of a cell is selected from the group consisting of the cell genome or extrachromosomal DNA further selected from the group consisting of an episome or a plasmid.

17. The nucleic acid fragment of claim 1 wherein at least one of the inverted repeats comprises SEQ ID NO:4 or SEQ ID NO:5.

18. The nucleic acid fragment of claim 1 wherein the amino acid sequence of the SB protein has at least an 80% amino acid identity to SEQ ID NO:1.

19. The nucleic acid fragment of claim 1 wherein at least one of the inverted repeats comprises at least one direct repeat and wherein the at least one direct repeat sequence comprises SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8 or SEQ ID NO:9.

20. The nucleic acid fragment of claim 1 wherein the direct repeat has a consensus sequence of SEQ ID NO:10.

21. The nucleic acid fragment of claim 1 wherein the direct repeat has at least an 80% nucleic acid sequence identity to SEQ ID NO:10.

22. A gene transfer system to introduce DNA into the DNA of a cell comprising:

a nucleic acid fragment comprising a nucleic acid sequence positioned between at least two inverted repeats wherein the inverted repeats can bind to an SB protein and wherein the nucleic acid fragment is capable of integrating into DNA of a cell; and

a transposase or nucleic acid encoding a transposase, wherein the transposase is an SB protein with an amino acid sequence sharing at least an 80% identity to SEQ ID NO:1.

23. The gene transfer system of claim 22 wherein the SB protein comprises SEQ ID NO:1.

24. The gene transfer system of claim 22 wherein the DNA encoding the transposase can hybridize to SEQ ID NO:1 under the following hybridization and wash conditions: in 30 % (v/v) formamide in 0.5xSSC, 0.1% (w/v) SDS at 42°C for 7 hours.

25. The gene transfer system of claim 22 wherein the transposase is provided to the cell as a protein.

26. The gene transfer system of claim 22 wherein the transposase is provided to the cell as nucleic acid encoding a transposase

27. The gene transfer system of claim 26 wherein the nucleic acid encoding a transposase is RNA.

28. The gene transfer system of claim 22 wherein the nucleic acid encoding the transposase is integrated into the genome of the cell.

29. The gene transfer system of claim 22 wherein the nucleic acid fragment is part of a plasmid or a recombinant viral vector.

30. The gene transfer system of claim 22 wherein the nucleic acid sequence comprises at least a portion of an open reading frame.

31. The gene transfer system of claim 22 wherein the nucleic acid sequence comprises at least a regulatory region of a gene.

32. The gene transfer system of claim 31 wherein the regulatory region is a transcriptional regulatory region.

33. The gene transfer system of claim 31 wherein the regulatory region is selected from the group consisting of a promoter, an enhancer, a silencer, a locus-control region, and a border element.

34. The gene transfer system of claim 22 wherein the cell is obtained from an animal.

35. The gene transfer system of claim 22 wherein the nucleic acid sequence comprises a promoter operably linked to at least a portion of an open reading frame.

36. The gene transfer system of claim 34 wherein the cell is a vertebrate or an invertebrate cell.

37. The gene transfer system of claim 36 wherein the invertebrate is obtained from a crustacean or a mollusk.

38. The gene transfer system of claim 36 wherein the cell is obtained from a fish or a bird.

39. The gene transfer system of claim 36 wherein the vertebrate is a mammal.

40. The gene transfer system of claim 39 wherein the cell is obtained from the group consisting of rodents, ungulates, sheep, swine and humans.

41. The gene transfer system of claim 22 wherein the DNA of a cell is selected from the group consisting of the cell genome or extrachromosal DNA.

42. The gene transfer system of claim 22 wherein at least one of the inverted repeats comprises SEQ ID NO:4 or SEQ ID NO:5.

43. The gene transfer system of claim 22 wherein the amino acid sequence of the SB protein has at least a 80% identity to SEQ ID NO:1.

44. The gene transfer system of claim 22 wherein at least one of the inverted repeats comprises at least one direct repeat and wherein the at least one direct repeat sequence comprises SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8 or SEQ ID NO:9.

45. The gene transfer system of claim 22 wherein the direct repeat has a consensus sequence of SEQ ID NO:10.

46. The gene transfer system of claim 22 wherein the nucleic acid sequence is part of a library of recombinant sequences.

47. The gene transfer system of claim 22 wherein the nucleic acid sequence is introduced into the cell using a method selected from the group consisting of:

particle bombardment;

electroporation;

microinjection;

combining the nucleic acid fragment with lipid-containing vesicles or DNA condensing reagents; and

incorporating the nucleic acid fragment into a viral vector and contacting the viral vector with the cell.

48. Nucleic acid encoding an SB protein, wherein the nucleic acid encodes a protein comprising SEQ ID NO:1 or a protein comprising an amino acid sequence with at least 80% identity to SEQ ID NO:1.

49. The nucleic acid of claim 48 in a nucleic acid vector.

50. The nucleic acid of claim 49 wherein the vector is a gene expression vector.

51. The nucleic acid of claim 50 wherein the vector is a plasmid.

52. The nucleic acid of claim 49 wherein the nucleic acid is a linear nucleic acid fragment.

53. Cells containing the nucleic acid of claim 48.

54. The nucleic acid of claim 53 wherein the cell is obtained from an animal.

55. The nucleic acid of claim 54 wherein the cell is obtained from a vertebrate or an invertebrate.

56. The nucleic acid of claim 55 wherein the vertebrate is a fish.

57. The nucleic acid of claim 55 wherein the vertebrate is a mammal.

58. The nucleic acid of claim 53 wherein the cell is an oocyte or an egg.

59. The nucleic acid of claim 53 wherein the cell is part of a tissue or organ.

60. The nucleic acid of claim 53 wherein the cell comprises one or more cells of an embryo.

61. The nucleic acid of claim 48 integrated in the genome of a cell.

62. An SB protein comprising the amino acid sequence of SEQ ID NO:1.

63. A method for producing a transgenic animal comprising the steps of:

introducing a nucleic acid fragment and a transposase into a pluripotent or totipotent cell wherein the nucleic acid fragment comprises a nucleic acid sequence positioned between at least two inverted repeats, wherein the inverted repeats can bind to an SB protein and wherein the nucleic acid fragment is capable of integrating into DNA in a cell and wherein the transposase is an SB protein having an amino acid sequence identity of least 80% to SEQ ID NO:1; and

growing the cell into an animal.

64. The method of claim 63 wherein the pluripotent or totipotent cell is selected from the group consisting of an oocyte, a cell of an embryo, an egg and a stem cell .

65. The method of claim 63 wherein the introducing step comprises a method selected from the group consisting of:

microinjection;

electroporation;

combining the nucleic acid fragment with cationic lipid vesicles or DNA condensing reagents; and

incorporating the nucleic acid fragment into a viral vector and contacting the viral vector with the cell.

66. The method of claim 65 wherein the viral vector is selected from the group consisting of a retroviral vector, an adenovirus vector or an adeno-associated viral vector, or a herpes virus.

67. The method of claim 63 wherein the animal is a mouse, a fish, an ungulate, a bird, or a sheep.

68. A method for introducing nucleic acid into DNA in a cell comprising the step of:

introducing a nucleic acid fragment comprising a nucleic acid sequence positioned between at least two inverted repeats into a cell wherein the inverted repeats can bind to an SB protein and wherein the nucleic acid fragment is capable of integrating into DNA in a cell in the presence of an SB protein.

69. The method of claim 68 wherein the method further comprises introducing an SB protein into the cell.

70. The method of claim 68 wherein the SB protein has an amino acid sequence comprising at least a 80% identity to SEQ ID NO:1.

71. The method of claim 69 wherein the SB protein is introduced to the cell as RNA.

72. The method of claim 68 wherein the cell comprises nucleic acid encoding an SB protein.

73. The method of claim 72 wherein the nucleic acid encoding the SB protein is integrated into the cell genome.

74. The method of claim 72 wherein the SB protein is stably expressed in the cell.

75. The method of claim 72 wherein the SB protein is under the control of an inducible promoter.

76. The method of claim 68 wherein the introducing step comprises a method for introducing nucleic acid into a cell selected from the group consisting of:

microinjection;

electroporation;

combining the nucleic acid fragment with cationic lipid vesicles or DNA condensing reagents; and

incorporating the nucleic acid fragment into a viral vector and contacting the viral vector with the cell.

77. The method of claim 76 wherein the viral vector is selected from the group consisting of a retroviral vector, an adenovirus vector or an adeno-associated viral vector.

78. The method of claim 68 wherein the method further comprises the step of introducing an SB protein or RNA encoding an SB protein into the cell.

79. The method of claim 68 wherein the cell is a pluripotent or a totipotent cell.

80. Transgenic animals produced by the method of claim 79.

81. The method of claim 68 wherein the nucleic acid sequence encodes a protein.

82. The method of claim 68 wherein the protein is a marker protein.

83. Cells producing the protein of claim 81.

84. Transgenic animals producing the recombinant protein produced by the method of claim 81.

85. A protein comprising the following characteristics:

an ability to catalyze the integration of nucleic acid into DNA of a cell;

capable of binding to the inverted repeat sequence of SEQ ID NOS:4 or 5; and

80% amino acid sequence identity to SEQ ID NO:1.

86. A protein comprising the following characteristics:

transposase activity;

a molecular weight range of about 35 kD to about 40 kD on about a 10% SDS-polyacrylamide gel; and

an NLS sequence, a DNA binding domain and a catalytic domain wherein the protein has at least about five-fold improvement in the rate for introducing a nucleic acid fragment into the nucleic acid of a cell as compared to the level obtained by non-homologous recombination.

87. A method for mobilizing a nucleic acid sequence in a cell comprising the steps of:

introducing the protein of claims 84 or 85 into a cell housing DNA containing the nucleic acid fragment according to claim 1, wherein the protein mobilizes the nucleic acid fragment from a first position within the DNA of a cell to a second position within the DNA of the cell.

88. The method of claim 87 wherein the DNA of a cell is genomic DNA.

89. The method of claim 87 wherein the first position within the DNA of a cell is extrachromosomal DNA.

90. The method of claim 87 wherein the second position within the DNA of a cell is extrachromosomal DNA.

91. The method of claim 87 wherein the protein is introduced into the cell as nucleic acid..

92. A method for identifying a gene in a genome of a cell comprising the steps of:

introducing a nucleic acid fragment and an SB protein into a cell, wherein the nucleic acid fragment comprises a nucleic acid sequence positioned between at least two inverted repeats into a cell wherein the inverted repeats can bind to the SB protein and wherein the nucleic acid fragment is capable of integrating into DNA in a cell in the presence of the SB protein;

digesting the DNA of the cell with a restriction endonuclease capable of cleaving the nucleic acid sequence;

identifying the inverted repeat sequences;

sequencing the nucleic acid close to the inverted repeat sequences to obtain DNA sequence from an open reading frame; and

comparing the DNA sequence with sequence information in a computer database.

93. The method of claim 92 wherein the restriction endonuclease recognizes a 6-base recognition sequence.

94. The method of claim 93 wherein the digesting step further comprises cloning the digested fragments or PCR amplifying the digested fragments.

95. A stable transgenic vertebrate line comprising a gene operably linked to a promoter, wherein the gene and promoter are flanked by inverted repeats, and wherein the inverted repeats can bind to an SB protein.

96. The stable transgenic vertebrate of claim 95 wherein the SB protein comprises SEQ ID NO:1 or an amino acid sequence with at least 80% homology to SEQ ID NO:1.

97. The stable transgenic vertebrate of claim 96 wherein the vertebrate is a fish.

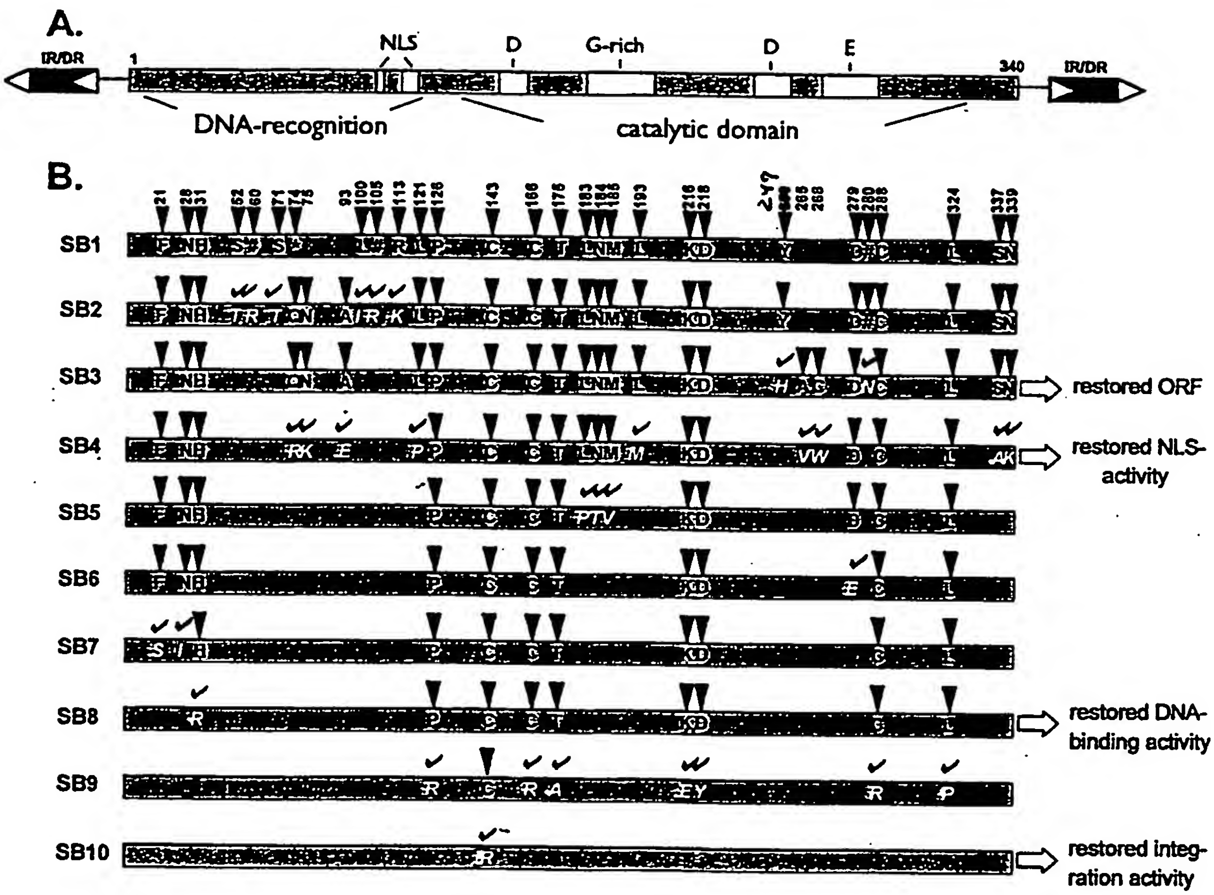
98. The stable transgenic vertebrate of claim 97 wherein the vertebrate is a zebrafish.

99. The stable transgenic vertebrate of claim 96 wherein the vertebrate is a mouse.

100. A protein with transposase activity that can bind to one or more of the following sequences:
SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, or SEQ ID NO:10.

101. An antibody capable of specifically binding to an SB protein.

FIG. 1A & 1B


A.
IR/DR
1
NLS
D
G-rich
D
E
340
IR/DR
DNA-recognition
catalytic domain
B.
SB1
SB2
SB3
SB4
SB5
SB6
SB7
SB8
SB9
SB10
restored ORF
restored NLS-activity
restored DNA-binding activity
restored integ-ration activity

```
1               ATGGGAAAA TCAAAAGAAA TCAGCCAAGA CCTCAGAAAA
                TACCCTTTT AGTTTTCTTT AGTCGGTTCT GGAGTCTTTT
-----------------------------------------------------------
51    AAAATTGTAG ACCTCCACAA GTCTGGTTCA TCCTTGGGAG CAATTTCCAA
      TTTTAACATC TGGAGGTGTT CAGACCAAGT AGGAACCCTC GTTAAAGGTT
-----------------------------------------------------------
101   ACGCCTGAAA GTACCACGTT CATCTGTACA AACAATAGTA CGCAAGTATA
      TGCGGACTTT CATGGTGCAA GTAGACATGT TTGTTATCAT GCGTTCATAT
151   AACACCATGG GACCACGCAG CCGTCATACC GCTCAGGAAG GAGACGCGTT
      TTGTGGTACC CTGGTGCGTC GGCAGTATGG CGAGTCCTTC CTCTGCGCAA
-----------------------------------------------------------
201   CTGTCTCCTA GAGATGAACG TACTTTGGTG CGAAAAGTGC AAATCAATCC
      GACAGAGGAT CTCTACTTGC ATGAAACCAC GCTTTTCACG TTTAGTTAGG
-----------------------------------------------------------
251   CAGAACAACA GCAAAGGACC TTGTGAAGAT GCTGGAGGAA ACAGGTACAA
      GTCTTGTTGT CGTTTCCTGG AACACTTCTA CGACCTCCTT TGTCCATGTT
-----------------------------------------------------------
301   AAGTATCTAT ATCCACAGTA AAACGAGTCC TATATCGACA TAACCTGAAA
      TTCATAGATA TAGGTGTCAT TTTGCTCAGG ATATAGCTGT ATTGGACTTT
-----------------------------------------------------------
351   GGCCGCTCAG CAAGGAAGAA GCCACTGCTC CAAAACCGAC ATAAGAAAGC
      CCGGCGAGTC GTTCCTTCTT CGGTGACGAG GTTTTGGCTG TATTCTTTCG
-----------------------------------------------------------
401   CAGACTACGG TTTGCAACTG CACATGGGGA CAAAGATCGT ACTTTTTGGA
      GTCTGATGCC AAACGTTGAC GTGTACCCCT GTTTCTAGCA TGAAAAACCT
451   GAAATGTCCT CTGGTCTGAT GAAACAAAAA TAGAACTGTT TGGCCATAAT
      CTTTACAGGA GACCAGACTA CTTTGTTTTT ATCTTGACAA ACCGGTATTA
-----------------------------------------------------------
501   GACCATCGTT ATGTTTGGAG GAAGAAGGGG GAGGCTTGCA AGCCGAAGAA
      CTGGTAGCAA TACAAACCTC CTTCTTCCCC CTCCGAACGT TCGGCTTCTT
-----------------------------------------------------------
551   CACCATCCCA ACCGTGAAGC ACGGGGGTGG CAGCATCATG TTGTGGGGGT
      GTGGTAGGGT TGGCACTTCG TGCCCCCACC GTCGTAGTAC AACACCCCCA
601   GCTTTGCTGC AGGAGGGACT GGTGCACTTC ACAAAATAGA TGGCATCATG
      CGAAACGACG TCCTCCCTGA CCACGTGAAG TGTTTTATCT ACCGTAGTAC
-----------------------------------------------------------
651   AGGAAGGAAA ATTATGTGGA TATATTGAAG CAACATCTCA AGACATCAGT
      TCCTTCCTTT TAATACACCT ATATAACTTC GTTGTAGAGT TCTGTAGTCA
701   CAGGAAGTTA AAGCTTGGTC GCAAATGGGT CTTCCAAATG GACAATGACC
      GTCCTTCAAT TTCGAACCAG CGTTTACCCA GAAGGTTTAC CTGTTACTGG
751   CCAAGCATAC TTCCAAAGTT GTGGCAAAAT GGCTTAAGGA CAACAAAGTC
      GGTTCGTATG AAGGTTTCAA CACCGTTTTA CCGAATTCCT GTTGTTTCAG
801   AAGGTATTGG AGTGGCCATC ACAAAGCCCT GACCTCAATC CTATAGAAAA
      TTCCATAACC TCACCGGTAG TGTTTCGGGA CTGGAGTTAG GATATCTTTT
851   TTTGTGGGCA GAACTGAAAA AGCGTGTGCG AGCAAGGAGG CCTACAAACC
      AAACACCCGT CTTGACTTTT TCGCACACGC TCGTTCCTCC GGATGTTTGG
-----------------------------------------------------------
901   TGACTCAGTT ACACCAGCTC TGTCAGGAGG AATGGGCCAA AATTCACCCA
      ACTGAGTCAA TGTGGTCGAG ACAGTCCTCC TTACCCGGTT TTAAGTGGGT
951   ACTTATTGTG GGAAGCTTGT GGAAGGCTAC CCGAAACGTT TGACCCAAGT
      TGAATAACAC CCTTCGAACA CCTTCCGATG GGCTTTGCAA ACTGGGTTCA
1001  TAAACAATTT AAAGGCAATG CTACCAAATA CTAG
      ATTTGTTAAA TTTCCGTTAC GATGGTTTAT GATC
```

FIG. 2A

FIG. 2B

Paired-like domain with Leucine-zipper

```
1   MGKSKEISQD LRKKIVDLHK SGSSLGAISK RLKVPRSSVQ TIVRKYKHHG
51  TTQPSYRSGR RRVLSPRDER TLVRKVQINP RTTAKDLVKM LEETGTKVSI
                NLS
101 STVKRVLYRH NLKGRSARKK PLLQNRHKKA RLRFATAHGD KDRTFWRNVL
                                               Glycine-rich box
151 WSDETKIELF GHNDHRYVWR KKGEACKPKN TIPTVKHGGG SIMLWGCFAA
201 GGTGALHKID GIMRKENYVD ILKQHLKTSV RKLKLGRKWV FQMDNDPKHT
251 SKVVAKWLKD NKVKVLEWPS QSPDLNPIEN LWAELKKRVR ARRPTNLTQL
301 HQLCQEEWAK IHPTYCGKLV EGYPKRLTQV KQFKGNATKY *
```

DD(34)E box

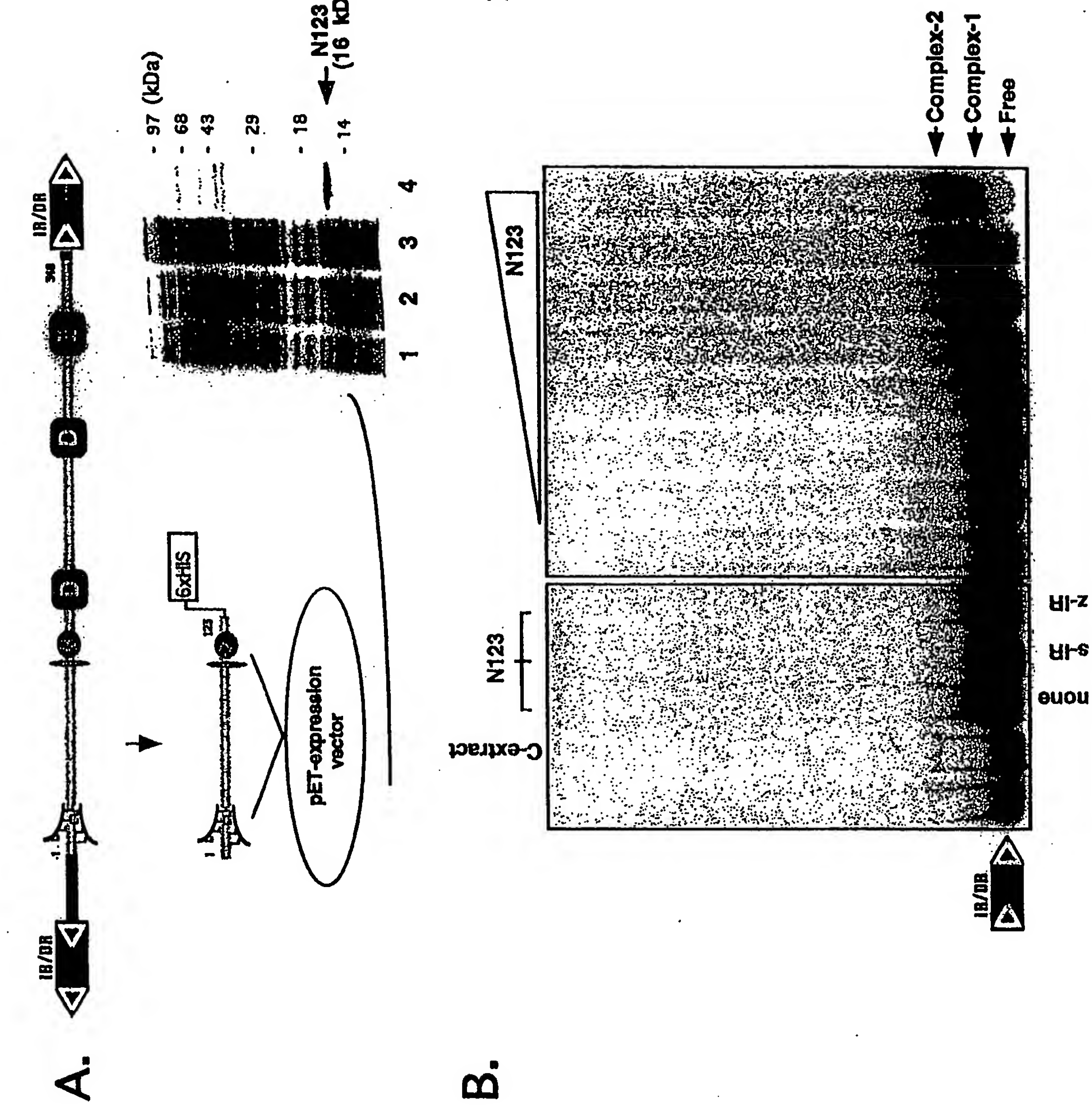
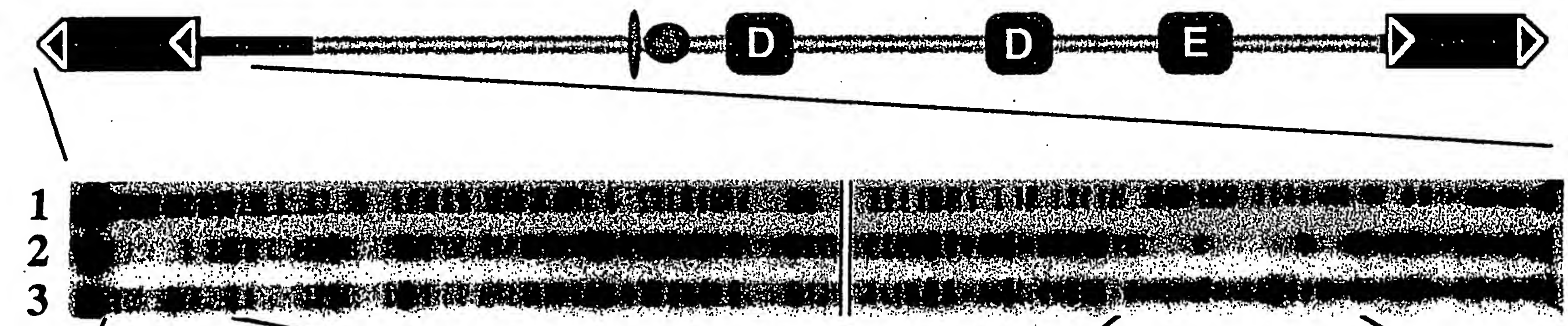

A.
IR/DR
6xHIS
pET-expression vector
97 (kDa)
68
43
29
18
14
N123 (16 kDA)
1 2 3 4
B.
C-extract
N123
none
s-IR
z-IR
Complex-2
Complex-1
Free

FIG. 3A+3B

FIGS. 4A-4C

A.



B.

```
gttgaagtcggaagtttacatacacttagg     - salmonid -     ccagtgggtcagaagtttacatacactaag
gtttaaaccagaagtttacacacactgtat     - zebrafish -    cttgaaagtc..aagtttacatacaataag
```

C.

```
External  tacagttgaagtcggaagtttacatacacttagg
Internal  tccagtgg..gtcagaagtttacatacactaagt
```

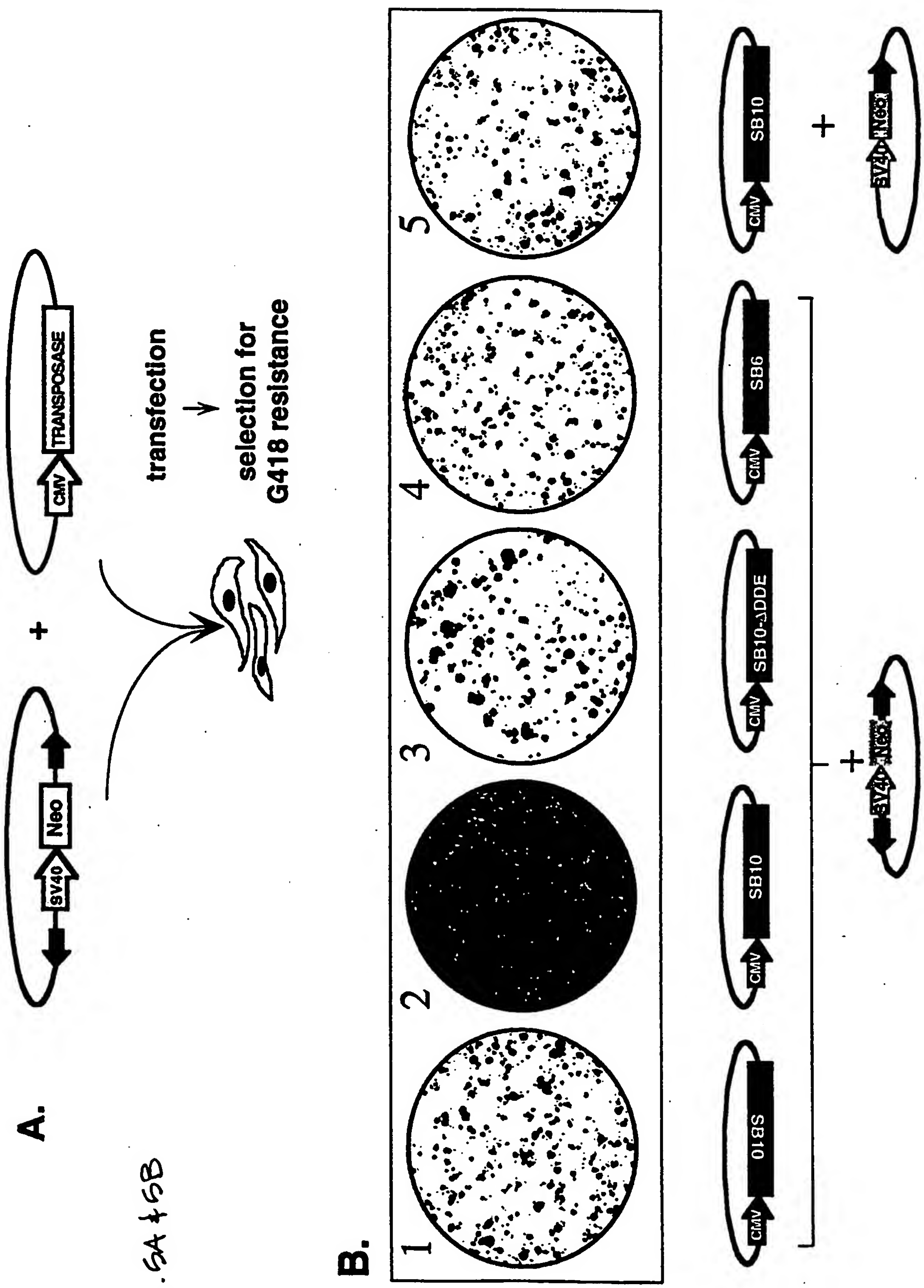

A.
SV40
Neo
+
CMV
TRANSPOSASE
transfection
selection for
G418 resistance
B.
1
2
3
4
5
CMV
SB10
SB10
SB10-ΔDDE
SB6
+
SV40
Neo

FIG. 5A & 5B

Fig. 6

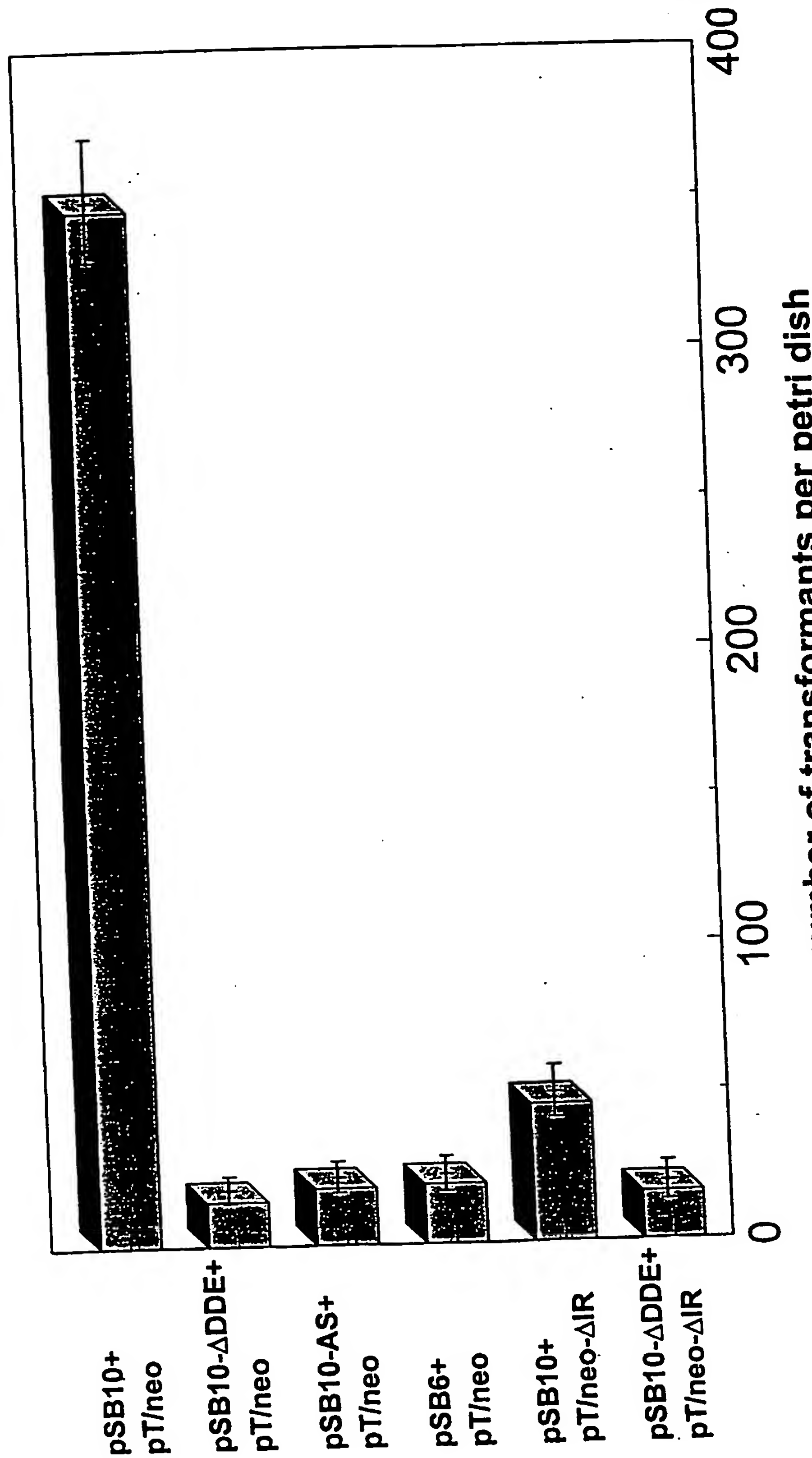

number of transformants per petri dish
0
100
200
300
400
pSB10+
pT/neo
pSB10-ΔDDE+
pT/neo
pSB10-AS+
pT/neo
pSB6+
pT/neo
pSB10+
pT/neo-ΔIR
pSB10-ΔDDE+
pT/neo-ΔIR

FIG. 7A

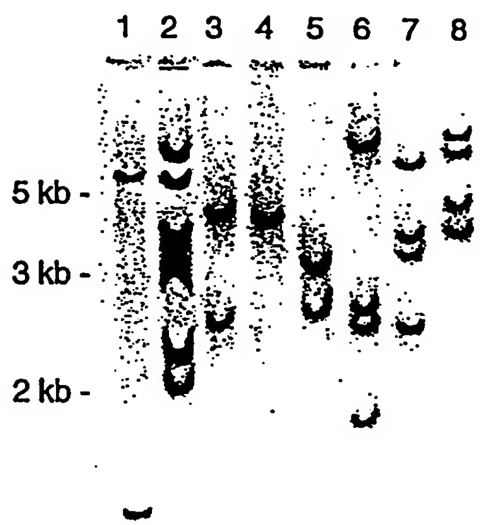
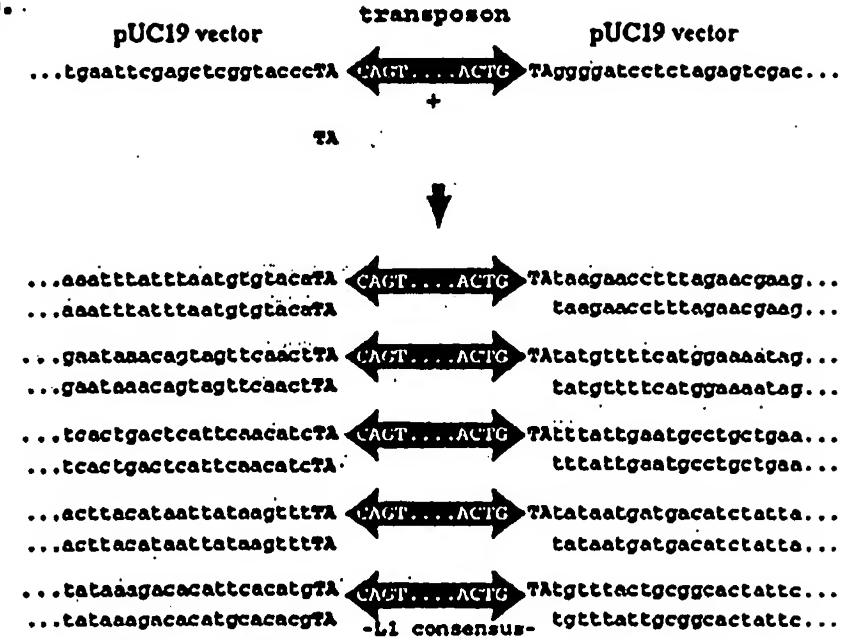


FIG. 7B

B.


pUC19 vector
transposon
pUC19 vector
...tgaattcgagctcggtacccTA
CAGT....ACTG
TAggggatcctctagagtcgac...
+
TA
...aaatttatttaatgtgtacaTA
TAtaagaacctttagaacgaag...
...aaatttatttaatgtgtacaTA
taagaacctttagaacgaag...
...gaataaacagtagttcaactTA
TAtatgttttcatggaaaatag...
...gaataaacagtagttcaactTA
tatgttttcatggaaaatag...
...tcactgactcattcaacatcTA
TAtttattgaatgcctgctgaa...
...tcactgactcattcaacatcTA
tttattgaatgcctgctgaa...
...acttacataattataagtttTA
TAtataatgatgacatctatta...
...acttacataattataagtttTA
tataatgatgacatctatta...
...tataaagacacattcacatgTA
TAtgtttactgcggcactattc...
...tataaagacacatgcacacgTA
-L1 consensus-
tgtttattgcggcactattc...

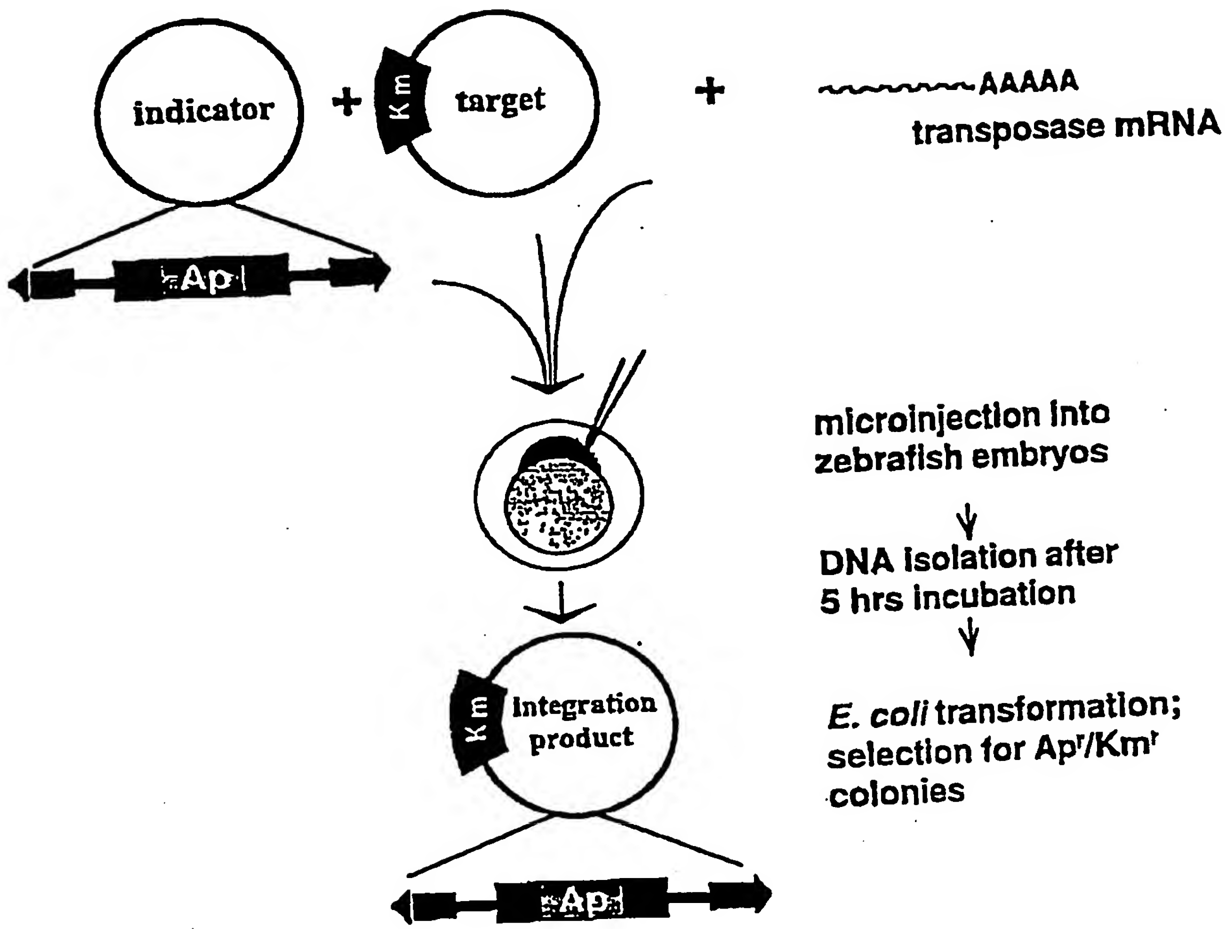

indicator
+
Km
target
+
AAAAA
transposase mRNA
Ap
microinjection into
zebrafish embryos
DNA isolation after
5 hrs incubation
E. coli transformation;
selection for Ap^r/Km^r
colonies
Km
Integration
product
Ap

FIG. 8

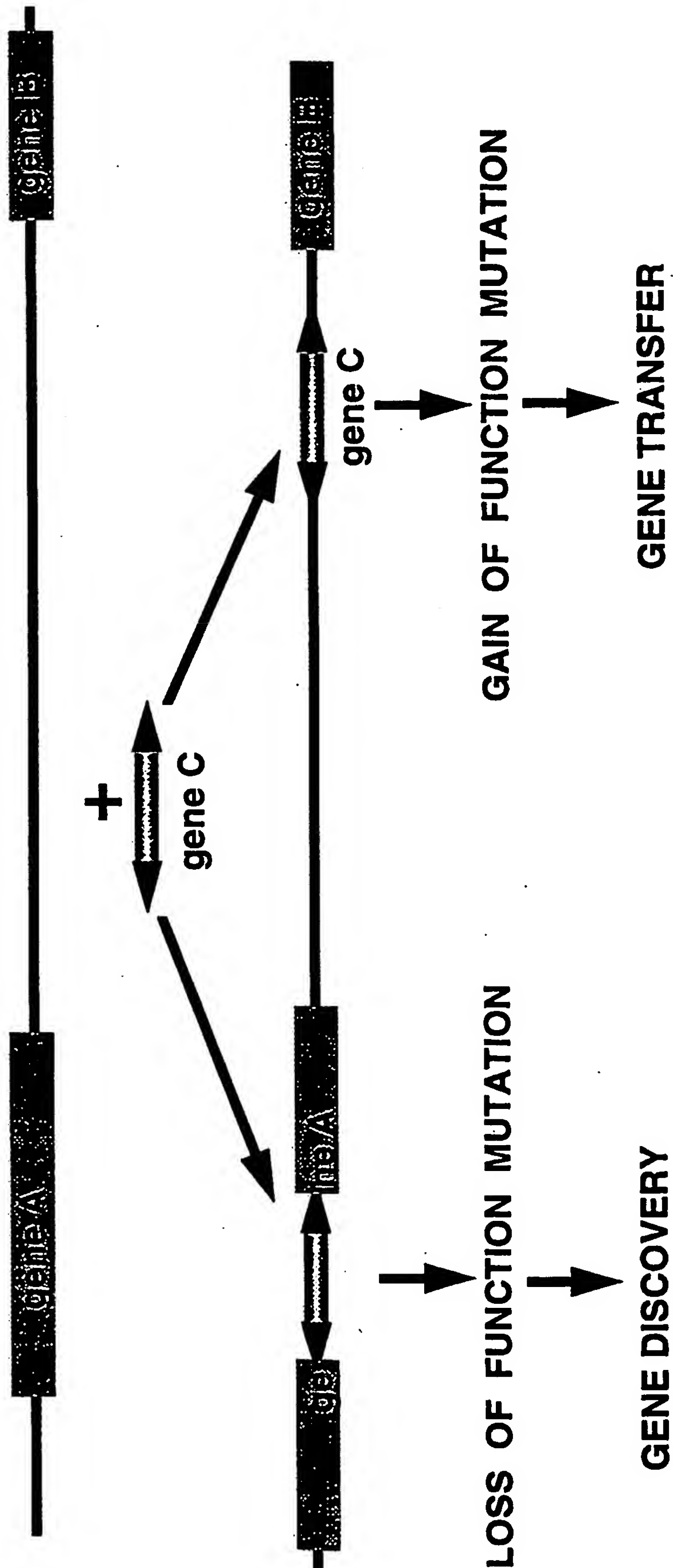
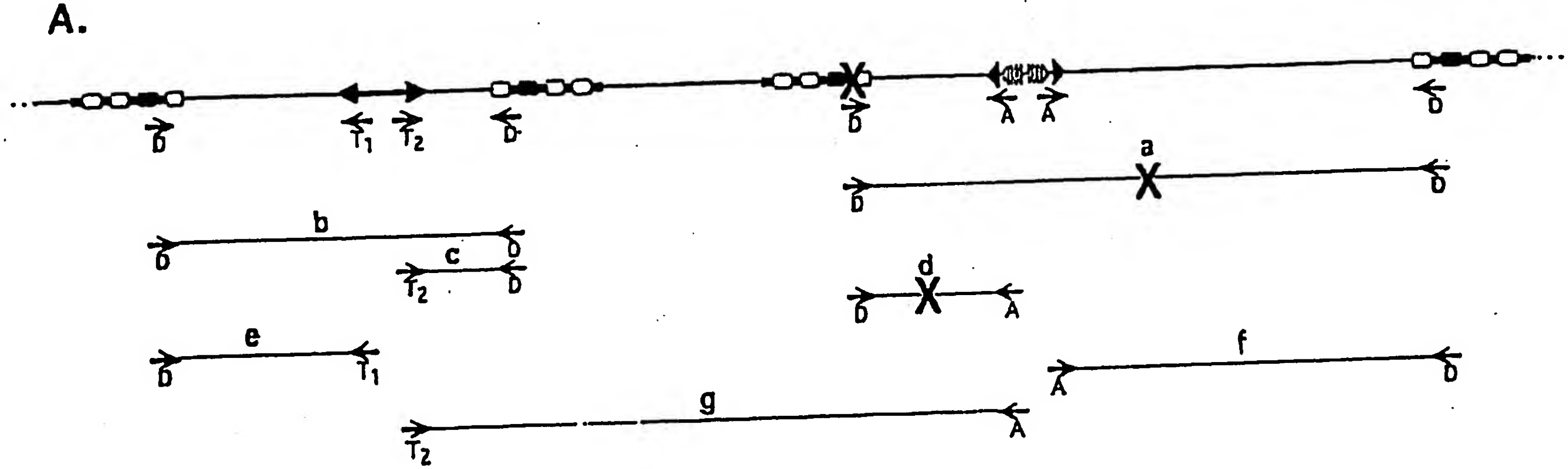
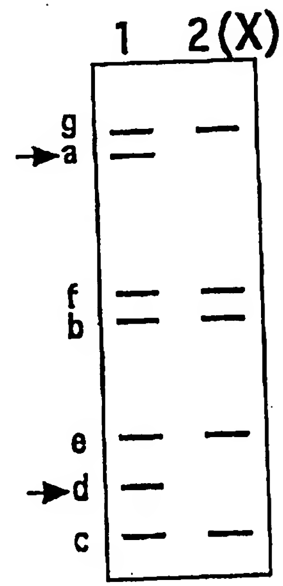

gene B
gene A
gene C
gene C
gene B
gene A
GAIN OF FUNCTION MUTATION
GENE TRANSFER
LOSS OF FUNCTION MUTATION
GENE DISCOVERY

FIG. 9

A.



B.



Generate sequence-tagged sites (STS) by isolation of fragments a and d and place them on a genetic map.

FIGS. 10A & 10B

INTERNATIONAL SEARCH REPORT

International ication No
PCT/US 98/04687

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12N15/90 C12N5/16 A01K67/027 C07K16/18

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12N A01K C07K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	IZSVAK Z ET AL: "Characterization of a Tc1-like transposable element in zebrafish (Danio rerio)." MOL GEN GENET, MAY 10 1995, 247 (3) P312-22, GERMANY, XP002072765 see the whole document ---	1-21
P,X	IVICS Z ET AL: "Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells." CELL, (1997 NOV 14) 91 (4) 501-10., XP002072766 see the whole document ---	1-101

-/--

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 July 1998	12/08/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Chakravarty, A

1

INTERNATIONAL SEARCH REPORT

International .lication No
PCT/US 98/04687

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	IVICS Z ET AL: "Identification of functional domains and evolution of Tc1-like transposable elements." PROC NATL ACAD SCI U S A, MAY 14 1996, 93 (10) P5008-13, UNITED STATES, XP002072767 see the whole document	

1

9/13

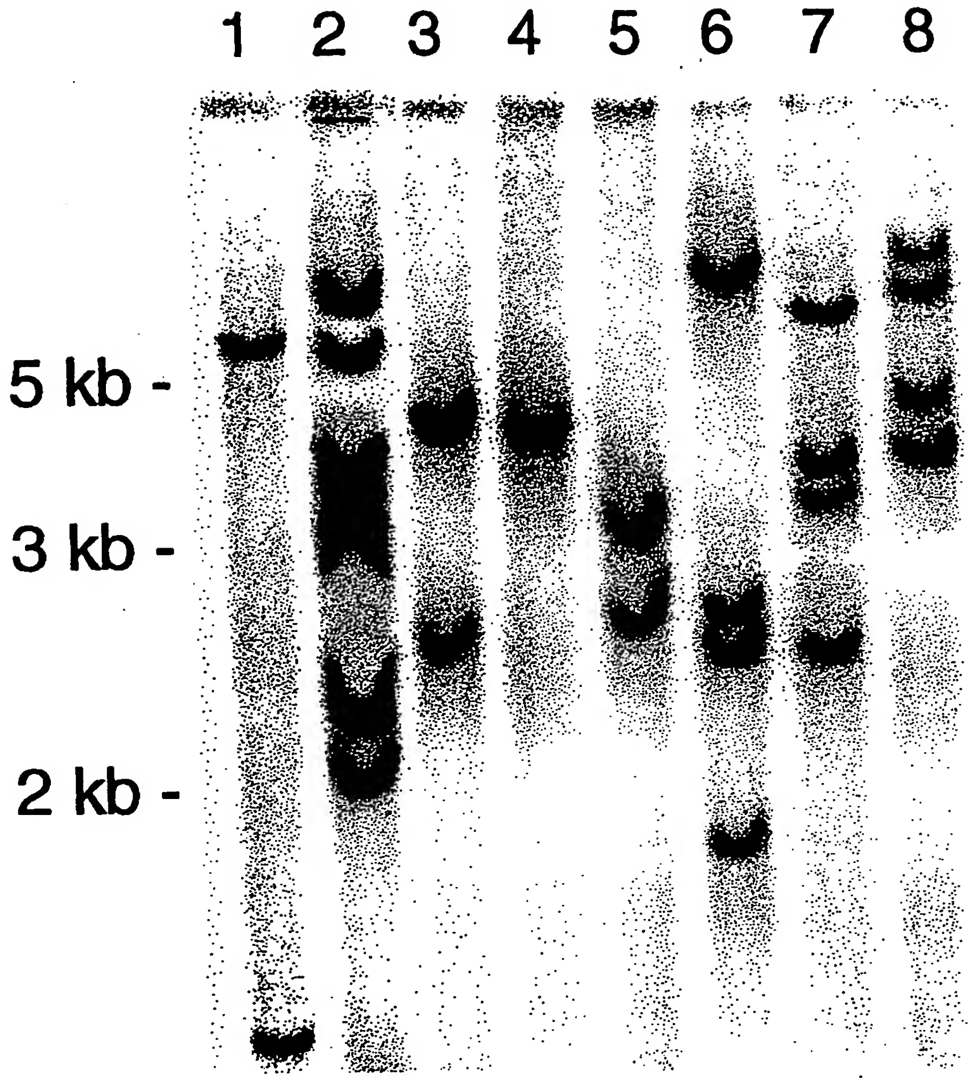
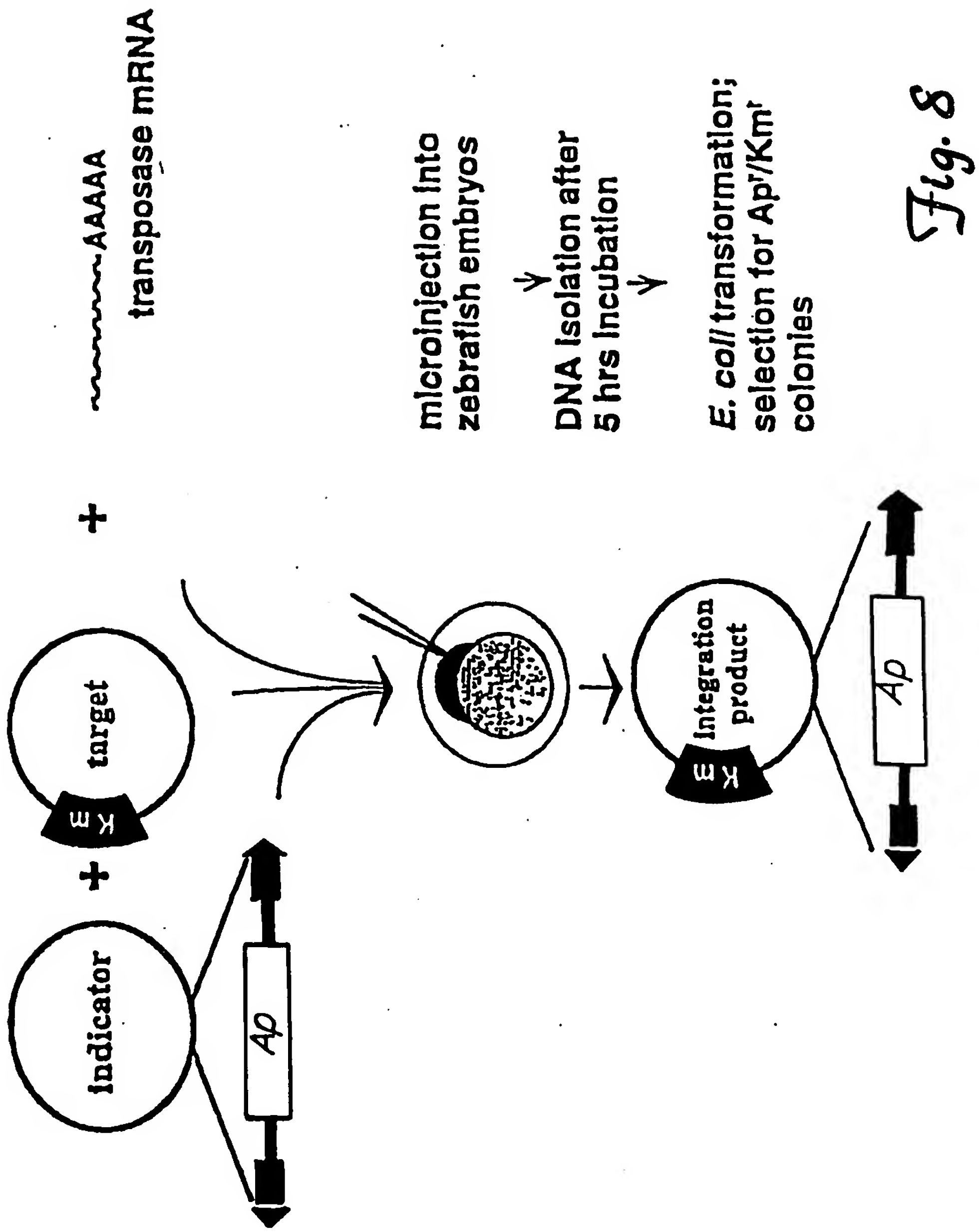


Fig. 7A

pUC19 vector	transposon	pUC19 vector
...tgaattcgagctcggtacccTA (SEQ ID NO:43)	CAGT ....ACTG	TAggggatcctctagagtcgac.... (SEQ ID NO:44)
	+	
TA	⇩	
(SEQ ID NO:45) ...aaatttatttaatgtgtacaTA	CAGT ....ACTG	TAtaagaacctttagaacgaag... (SEQ ID NO:46)
47...aaatttatttaatgtgtacaTA		taagaacctttagaacgaag...48
49...gaataaacagtagttcaactTA	CAGT ....ACTG	TAtatgttttcatggaaaatag...50
51...gaataaacagtagttcaactTA		tatgttttcatggaaaatag...52
53...tcactgactcattcaacatcTA	CAGT ....ACTG	TAtttattgaatgcctgctgaa...54
55...tcactgactcattcaacatcTA		tttattgaatgcctgctgaa...56
57...acttacataattataagtttTA	CAGT ....ACTG	TAtataatgatgacatctacta...58
59...acttacataattataagtttTA		tataatgatgacatctatta...60
61...tataaagacacattcacatgTA	CAGT ....ACTG	TAtgtttactgcggcactattc...62
63...tataaagacacatgcacacgTA		tgtttattgcggcactattc...2
	-L1 CONSENSUS	

Fig. 7B



Indicator
Ap
target
Km
AAAAA
transposase mRNA
microinjection into zebrafish embryos
DNA isolation after 5 hrs incubation
E. coli transformation; selection for Ap^r/Km^r colonies
Integration product
Km
Ap

Fig. 8

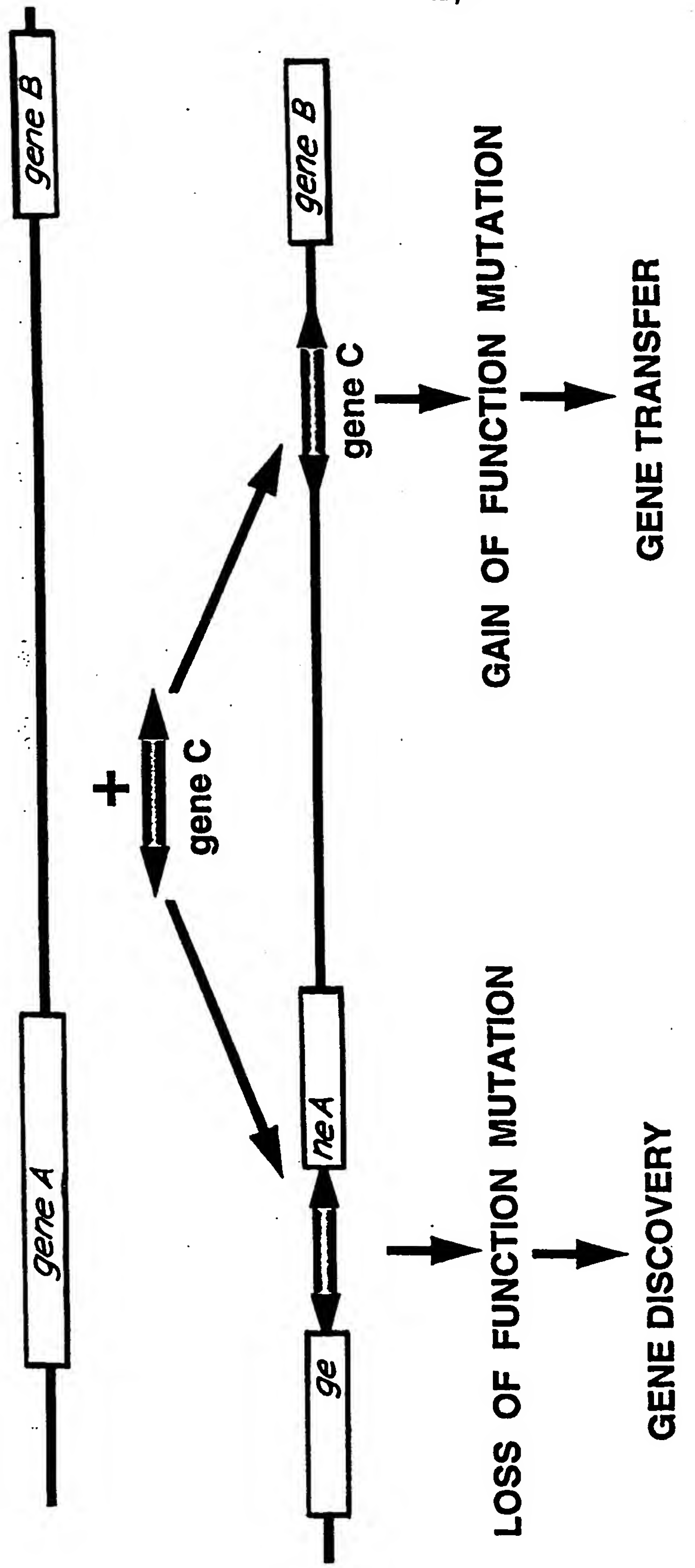

gene B
gene A
+
gene C
gene B
gene C
neA
ge
GAIN OF FUNCTION MUTATION
GENE TRANSFER
LOSS OF FUNCTION MUTATION
GENE DISCOVERY

Fig. 9

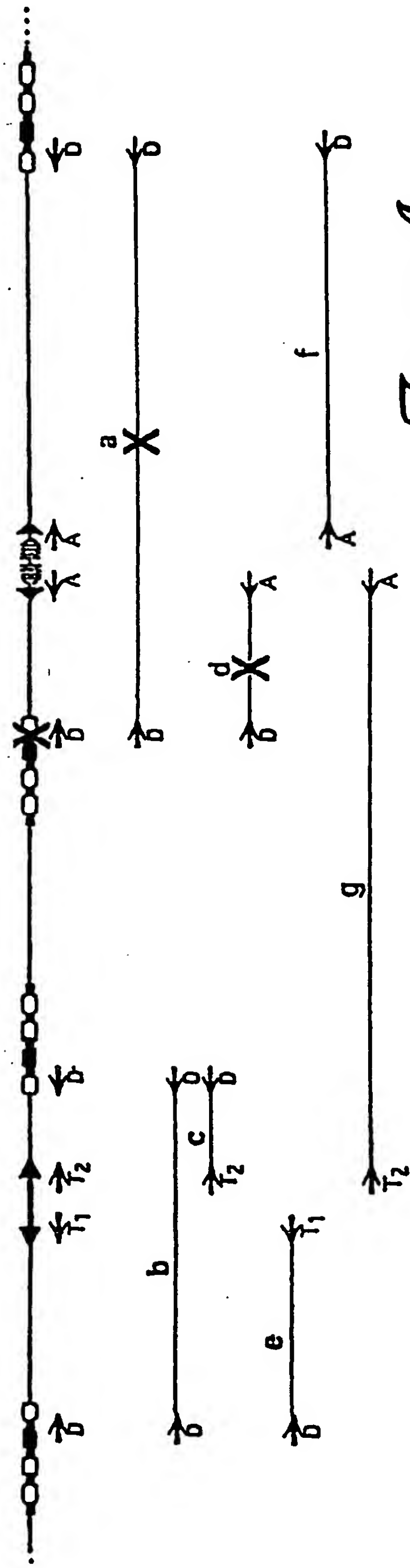


Fig. 10A

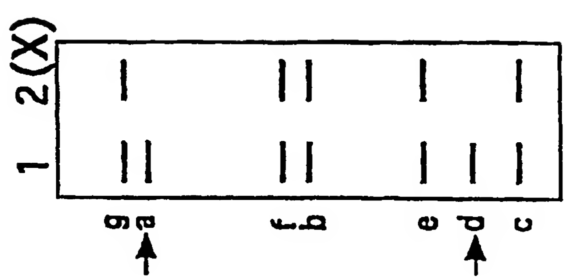


Fig. 10B

Generate sequence-tagged sites (STS) by isolation of fragments a and d and place them on a genetic map.

INTERNATIONAL SEARCH REPORT

International ication No
PCT/US 98/04687

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12N15/90 C12N5/16 A01K67/027 C07K16/18

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12N A01K C07K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	IZSVAK Z ET AL: "Characterization of a Tc1-like transposable element in zebrafish (Danio rerio)." MOL GEN GENET, MAY 10 1995, 247 (3) P312-22, GERMANY, XP002072765 see the whole document	1-21
P,X	IVICS Z ET AL: "Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells." CELL, (1997 NOV 14) 91 (4) 501-10., XP002072766 see the whole document	1-101

-/--

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 July 1998	12/08/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Chakravarty, A

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International .ication No
PCT/US 98/04687

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropnate, of the relevant passages	Relevant to claim No.
A	IVICS Z ET AL: "Identification of functional domains and evolution of Tc1-like transposable elements." PROC NATL ACAD SCI U S A, MAY 14 1996, 93 (10) P5008-13, UNITED STATES, XP002072767 see the whole document	

1

Form PCT/ISA/210 (continuation of second sheet) (July 1992)